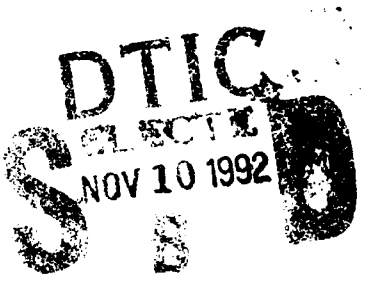
Navy Personnel Research and Development Center
San Diego, California 92152-6800 TR-93-2 October 1992

(2)

Detecting and Measuring Improvements in Validity

Larry V. Hedges
Betsy Jane Becker
John H. Wolfe



NPRDC-TR-93-2 October 1992

Detecting and Measuring Improvements in Validity

Larry V. Hedges
University of Chicago

Betsy Jane Becker
Michigan State University

John H. Wolfe
Navy Personnel Research and Development Center

Reviewed by
Frank L. Vicino

Approved by
W. A. Sands
Director, Testing Systems Department

Released by
Thomas F. Finley
Captain, U.S. Navy
Commanding Officer
and
Richard C. Sorenson
Technical Director (Acting)

Navy Personnel Research and Development Center
San Diego, California 92152-6800

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATE COVERED
	October 1992	Final--18 Sep 89-30 Sep 90

4. TITLE AND SUBTITLE
Detecting and Measuring Improvements in Validity

5. FUNDING NUMBERS
Reimbursable: Operations & Maintenance, Army
Work Unit No.: MIPR89-R-114
Contract: DAAL03-86-D-0001
John H. Wolfe (COTR)

6. AUTHOR(S)
Larry V. Hedges, Betsy Jane Becker, John H. Wolfe

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Battelle Memorial Institute
505 King Avenue
Columbus, OH 43201

8. PERFORMING ORGANIZATION REPORT NUMBER
NPRDC-TR-93-2

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Navy Personnel Research and Development Center
San Diego, CA 92152-6800

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Functional Area: Personnel Systems
Product Line: Computerized Testing
Effort: New Measures of Ability

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The Navy Personnel Research and Development Center (NPRDC) has developed new aptitude tests for possible addition to the Armed Services Vocational Aptitude Battery (ASVAB). Validity studies are currently underway to determine whether the new tests produce an increment in validity using either a training or a job performance criterion. Because most samples are small, it is necessary to pool information across sites to obtain a sufficiently powerful test for incremental validity.

Methods for obtaining the sampling distributions of incremental validities (the differences between multiple correlations) from the same and from independent samples were developed and are presented. These results were applied to yield methods for pooling incremental validities, testing the statistical significance of pooled validities, constructing confidence intervals for the pooled incremental validity, and conducting a power analysis of the pooled test for incremental validity. It was concluded that the test for the statistical significance of the pooled estimate should have adequate power to detect increments in validity of .02 given pooled sample sizes of $N \geq 4{,}000$.

14. SUBJECT TERMS
Meta-analysis, test validity, psychometrics, personnel psychology

15. NUMBER OF PAGES
67

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

FOREWORD

The Navy Personnel Research and Development Center is the lead laboratory for the Enhanced Computerized Aptitude Testing (ECAT) project. The purpose of the project is to assess the cost/benefits of adding new aptitude tests to the Armed Services Vocational Aptitude Battery (ASVAB). This report solves the important problem of how to combine the results from different studies with different criteria in order to arrive at estimates of the incremental validity of adding new tests to the ASVAB. The issue is of practical importance because many of the samples under investigation are too small to allow firm conclusions to be drawn unless their data are combined with those of other samples. This report will be useful both to military personnel researchers and to a broad civilian research community concerned with the validity of aptitude tests.

This effort was conducted under the ECAT project sponsored by the Office of the Assistant Secretary of Defense (Force Management & Personnel, Military Manpower & Personnel Policy). It was funded by Headquarters, U.S. Military Entrance Processing Command (USMEPCOM) with U. S. Army Operations and Maintenance funds (MIPR 89-R-114). The report was written under the Army Research Office contract DAAL03-86-D-0001, TCN 89-517, D.O. 1723 with Battelle Memorial Institute.

John H. Wolfe was the Contracting Officer's Technical Representative (COTR) for the task.

THOMAS F. FINLEY
Captain, U. S. Navy
Commanding Officer

RICHARD C. SORENSON
Technical Director (Acting)

SUMMARY

Problem

The Navy Personnel Research and Development Center (NPRDC) has developed new aptitude tests for possible addition to the Armed Services Vocational Aptitude Battery (ASVAB). A computerized version of the ASVAB (the CAT-ASVAB) was also developed. Validity studies are currently underway to determine whether the new tests on the CAT-ASVAB produce an increment in validity computed using a job performance criterion. However, few single sites (schools) have a large enough sample to produce a sufficiently powerful test for the validity increment expected. Consequently, it will be necessary to pool information across sites to obtain a sufficiently powerful test for incremental validity. The methods for such pooling had not previously been developed.

Objective

The objective of this research was to develop statistical methods for pooling estimates of incremental validity across independent studies (sites), estimate the standard errors and a confidence interval for the pooled incremental validity, and test the statistical significance of the incremental validity.

Approach

A search of the literature on combining statistical estimates was conducted to determine applicable methods. The statistical literature on the sampling theory of multiple correlations was also searched. Mathematical (analytic) methods were used to derive procedures that were not previously available.

Results

Methods for obtaining the sampling distributions of incremental validities (the differences between multiple correlations) from the same and from independent samples were developed. These results were applied to yield methods for pooling incremental validities, testing the statistical significance of pooled validities, and constructing confidence intervals for the pooled incremental validity. They were also applied to a power analysis of the pooled test for incremental validity.

Conclusion

Pooling estimates across sites provides a viable strategy for estimating the incremental validity. If a single sample is used in each site to assess incremental validity, the test for the statistical significance of the pooled estimate will have adequate power to detect increments in validity of .02 with pooled sample sizes of $N \geq 4,000$.

Recommendation

Estimates of the incremental validity of alternative test batteries should be based on pooled estimates derived from several samples, using the methods outlined in this report.

CONTENTS

LIST OF TABLES

INTRODUCTION

Problem

Current practice involves the use of a battery of tests to predict a criterion. We want to determine whether adding another battery of tests to the operational battery improves validity and by what amount. Because the increment in validity expected from the new tests is small (e.g., .02), a sample size of several thousand may be needed to detect the validity increment with high statistical power. The tests are used in a variety of sites (schools), but few single schools have sufficiently large enrollments to carry out a powerful study of the improvements in validity that might result from the addition of the battery of new tests.

Alternatively, current practice may involve the use of one test battery to predict a criterion and we may wish to know if an alternative test battery has greater predictive validity. For example, we may wish to compare the validity of a paper and pencil version of a test battery to that of a computerized adaptive version of the same test battery.

Objective

In these situations, pooling of information on incremental validity across several sites (e.g., schools) may provide a way to test the increment in validity with high statistical power and to estimate precisely the improvement in validity that results from the addition of the new tests or the use of the alternative tests. It is assumed that the criterion scores used in different sites are too dissimilar to permit combination of raw data. This is likely to be the case when the criteria are training grades, performance ratings, or simulations of work skills that are unique to the individual school site.

This report develops a method for combining estimates of incremental validity across sites to obtain the most precise estimate of the average incremental validity. It also provides procedures for estimating the standard error of the incremental validity, for establishing confidence intervals about the incremental validity, and for testing the combined significance of the validity-study results. Finally the report shows how to estimate the variance in incremental validity parameters and how to test its statistical significance.

Background

The Navy Personnel Research and Development Center (NPRDC) has been engaged in a project to evaluate new aptitude tests that measure abilities not covered in the existing battery of ten tests in the Armed Services Vocational Aptitude Battery (ASVAB). It has also been engaged in the development of a computerized adaptive version of the ASVAB called the CAT-ASVAB. Validity

studies are currently underway to determine the magnitude of incremental validity obtained by using the new tests as supplements to the ten operational ASVAB tests that are used as predictors of school and job performance. Navy studies should also provide data on the incremental validity of the CAT-ASVAB over the operational paper and pencil version of the test. The present report suggests methodology for carrying out studies of incremental validity in connection with these recent NPRDC developments.

INCREMENTAL VALIDITY

Given a single site, we might define the validity of test battery 1 as the multiple correlation R_1 of the a tests in that battery with the criterion. Define the validity of test battery 2 (which may consist of tests in battery 1 plus some new tests) as the multiple correlation R_2 of the tests in battery 2 with the criterion. It is helpful to distinguish the true or population values of validities from their sample estimates. Hence denote the sample estimates of the validities by R_1 and R_2 and denote the population values corresponding to these sample estimates by P_1 and P_2. The idea of incremental validity also arises in connection with the comparison of two alternative test batteries; for example, a paper and pencil test battery versus a computerized adaptive test battery. In this case, the two test batteries may not share any single test. In this case, the validities R_1 and P_1 are the sample and population multiple correlations of the a tests in battery 1 with the criterion. The sample and population validities of test battery 2, R_2 and P_2, respectively, are the sample and population multiple correlations of the b tests in battery 2 with the criterion.

The two multiple correlations that are used as validity coefficients are stochastically dependent when they are computed from measurements on the same sample of individuals. The correlations are stochastically independent when they are computed from independent samples.

Testing the Statistical Significance of Incremental Validity at a Single Site

At each site, the incremental validity study compares a sample validity R_1 with another sample validity R_2 to determine whether P_2 is larger than P_1. Formally this involves a test of the hypothesis that the population validity P_2 associated with R_2 exceeds the population validity P_1 associated with R_1; that is, a test of the hypothesis

$$H_0 : P_2 = P_1,$$

versus the alternative that $P_2 > P_1$.

Since P_1 and P_2 are nonnegative, $P_2 > P_1$ implies that $P_2^2 > P_1^2$; therefore a test for $P_2 > P_1$ is identical to a test for $P_2^2 > P_1^2$. The details of the hypothesis test depend on whether the same sample is used to compute both R_1 and R_2 or whether R_1 and R_2 are computed from independent samples.

Note that the artifacts of criterion unreliability and restriction of range do not alter the procedures for testing hypotheses about incremental validity.

R_1 and R_2 Computed from the Same Sample

Case 1. If R_1 and R_2 are computed from the same sample and if the predictors of R_1 are a subset of the predictors for R_2, then the appropriate test for incremental validity is the usual test for change in multiple correlation. Let $\underline{a}$ be the number of tests used as predictors in R_1, let $\underline{b} > \underline{a}$ be the number of tests used as predictors in R_2, and let n be the sample size. The test statistic is

$$F = \frac{(R_2^2 - R_1^2)(n - \underline{b} - 1)}{(1 - R_2^2)(\underline{b} - \underline{a})}, \qquad (1)$$

which is compared to the critical value for an F distribution with $(\underline{b} - \underline{a})$ and $(n - \underline{b} - 1)$ degrees of freedom.

Case 2. If R_1 and R_2 are computed from the same sample but their predictor sets are disjoint (e.g., when R_1 is computed from a pre-enlistment test battery and R_2 is computed from a post-enlistment test battery), the usual F-test for change in multiple correlation cannot be used. Let $\underline{a}$ be the number of predictors used to compute R_1 and $\underline{b}$ be the number of predictors used to compute R_2. Here $\underline{b}$ need not be larger than $\underline{a}$. A large sample test for the significance of the incremental validity uses the statistic

$$X^2 = \frac{(\hat{R}_2^2 - \hat{R}_1^2)^2}{\hat{\sigma}_\infty^2(\hat{d}^*)}, \qquad (2)$$

where $\hat{R}_2^2$ and $\hat{R}_1^2$ are the corrected squared multiple correlations for the two predictor sets and $\hat{\sigma}_\infty^2(\hat{d}^*)$ is the asymptotic variance of the difference in corrected squared multiple correlations. Much of the mathematical development in this paper will be devoted to estimating $\hat{\sigma}_\infty^2(\hat{d}^*)$. The test statistic X^2 (Equation 2) has a chi-square distribution with one degree of freedom when there is no incremental validity (but both of the validities are nonzero) and the sample size is large. The hypothesis of no incremental validity is rejected at significance level α if the computed value of X^2 exceeds the $100(1-\alpha)$ percentile point of the chi-square distribution with one degree of freedom.

R_1 and R_2 Computed from Independent Samples

If R_1 and R_2 are computed from independent samples (e.g., when R_1 is computed from the scores of subjects who took a paper and pencil test battery and R_2 is computed from the scores of subjects who took a computerized adaptive test battery), the usual F-test for change in multiple correlation cannot be used. Let $\underline{a}$ be the number of predictors used to compute R_1 and $\underline{b}$ be the number of predictors used to compute R_2. Here $\underline{b}$ need not be larger than $\underline{a}$. Let n_1 be the sample size on which R_1 is based and let n_2 be the sample

size on which R_2 is based. A large sample test for the significance of the incremental validity uses the statistic

$$X^2 = \frac{(R_2 - R_1)^2}{\frac{(1 - R_1^2)^2}{n_1} + \frac{(1 - R_2^2)^2}{n_2}} . \qquad (3)$$

The test statistic (Equation 3) has a chi-square distribution with one degree of freedom when there is no incremental validity (but both of the validities are nonzero) and both n_1 and n_2 are large. The hypothesis of no incremental validity is rejected at significance level α if the computed value of X^2 exceeds the $100(1-\alpha)$ percentile point of the chi-square distribution with one degree of freedom.

The design that involves computing R_1 and R_2 from the same sample yields more powerful tests for incremental validity. Consequently it is the design of choice wherever it is feasible.

Indices of Incremental Validity

Two indices of incremental validity might be computed. One index is simply the difference in validities. That is, the index, d, is the difference in (unsquared) multiple correlations. The sample value of this index is

$$d = R_2 - R_1$$

and the population value is

$$\delta = P_2 - P_1.$$

The index is conventionally used in personnel psychology and is, for example, used in the style of utility analyses described by Cronbach and Gleser (1965).

An alternative index of incremental validity is the R-squared change: the difference in squared multiple correlations. The sample value of this index is

$$d^* = R_2^2 - R_1^2$$

and the population value is

$$\delta^* = P_2^2 - P_1^2.$$

This index has the virtue that it is interpretable in terms of "additional variance accounted for" by the new test battery.

USING COMBINED SIGNIFICANCE TO STUDY INCREMENTAL VALIDITY

One of the oldest methodologies for combining the results of independent studies uses the observed significance levels (p's or probabilities) from series of significance tests. Tests of combined significance utilize the probability values from series of studies examining the same research question. Several combined significance methods have been outlined in the social-science literature by Rosenthal (1978) and previously by Mosteller and Bush (1954).

Although more than 15 distinct methods for summarizing observed probabilities have been proposed, the methods share some similarities. Table 1 lists the methods and shows that they fall into two major groups. One group comprises tests based on the fact that the observed significance levels are uniformly distributed under the null hypothesis of "no effect" in any study. The other methods involve transformations of the observed significance levels to other statistical variables (e.g., probabilities transformed to normal variates). All of the methods listed in Table 1 provide tests of the same null model for the series of studies. We outline that model for studies of incremental validity below.

Table 1
Methods for Summarizing Independent Significance Values

Methods Requiring Transformation of p Values

Indicator Function Methods

- Wilkinson method (Wilkinson, 1951)
- Tippett method (Tippett, 1931)
- Sign test
- Chi-square method

Inverse Probability Methods

Inverse Normal Distribution Methods

- Stouffer method (Stouffer et al., 1949)
- Weighted Stouffer method
- Mean z method

Inverse t Distribution Method

- Winer method (Winer, 1971)

Inverse Chi-square Distribution Methods

- Inverse chi-square method
- Weighted inverse chi-square method

Logistic Function Methods

- Fisher method (Fisher, 1932)
- Good (weighted Fisher) method
- Logit method

Methods not Requiring Transformation of p-values

- Sum of p's method
- Mean p method

Model for Study Results

Suppose that there are k independent validity studies, each yielding a test of incremental validity. We consider as illustration the situation in which a single set of subjects provides the information on incremental validity within every study (i.e., the estimates of incremental validity are d

and d^* as in our previous notation system). Each study in the series examines a test of the difference in validity between one battery including $\underline{a}$ tests and a second battery of $\underline{b}$ tests (where $\underline{b} > \underline{a}$). Each study might represent a training school or site for which it is useful to predict job performance.

As in the case in which parametric estimates of incremental validity are of interest, the ith study is assumed to provide a test of either δ_i or δ_i^*. In study i, $\delta_i = P_{i2} - P_{i1}$ is the difference in unsquared multiple correlations in the population, and $\delta_i^* = P_{i2}^2 - P_{i1}^2$ is the difference in squared multiple correlations. The null hypothesis (the model of no added validity) for the ith study would be either

$$H_0 : \delta_i = 0,$$

or equivalently $$H_0 : \delta_i^* = 0.$$

In each study, the usual F-test for R-squared change provides a significance test of the null hypothesis based on the sample estimates of incremental validity. The F statistic for incremental validity for a sample of size n from a single study or school is given by Equation 1 or

$$F = \frac{(R_2^2 - R_1^2)(n - \underline{b} - 1)}{(1 - R_2^2)(\underline{b} - \underline{a})} .$$

This statistic is distributed as a central F value with $(\underline{b} - \underline{a})$ and $(n - \underline{b} - 1)$ degrees of freedom under the null model of no contribution to validity from the added test battery in the study. The significance value from this test is the probability (p) of observing a value equal to F or larger in the F distribution with $(\underline{b} - \underline{a})$ and $(n - \underline{b} - 1)$ degrees of freedom.

From each F-test is obtained an observed upper-tail significance level. The observed probability from the ith study is p_i, and the data used in the combined significance tests are the k one-tailed probability values $p_1, p_2, \cdots, p_k$.

Null Hypothesis

The null hypothesis for tests of combined significance is an omnibus hypothesis, namely that the null hypothesis is true in <u>every one</u> of the studies in the synthesis. Thus the overall null model for any combined significance test for validity studies is

$$H_0: \quad \delta_1 = \delta_2 = \ldots = \delta_k = 0,$$

in the case of differences in multiple correlations, or equivalently

$$H_0: \quad \delta_1^* = \delta_2^* = \ldots = \delta_k^* = 0,$$

when differences in squared correlations represent added validity. Regardless of the parameter(s) used to represent validity, the null model for the series of studies is that the additional tests do <u>not</u> increase validity in any population studied.

One assumption of the combined significance methods is that the alternatives to the null hypothesis are one-sided. Typically this assumption appears as a restriction that the parameter tested in each study cannot be negative, and it leads to the condition that only one-tailed significance values are used in tests of combined significance. In some cases, this restriction requires redefining the parameters of the hypothesis for each study. For example, a hypothesis might be restated by defining the parameter of interest as θ^2 rather than θ, if both negative and positive θ values were interesting. In the case of validity studies, the alternative hypothesis is naturally one-sided, because additional tests can only increase the validity of prediction in each population, not decrease it.

Though the null hypothesis for the combined significance summaries is quite simple, deviations from the null hypothesis can occur in a variety of different ways. Thus, the interpretation of a rejected null hypothesis is not completely straightforward (see also Becker, 1987).

Let us consider the validity-study context. One way that the null model can be false is if <u>all</u> populations studied show increased validity because of the added tests. However, the null model is also false when a <u>single</u> population shows an increase in validity and the others do not. Both of these outcomes should lead to the rejection of the null hypothesis based on a test of combined significance, but they represent situations that are qualitatively very different.

Additionally, the combined significance tests themselves perform differently with regard to the detection of these various patterns of outcomes (i.e., different alternatives to the null model). Statistical theory (e.g., Oosterhoff, 1969) has shown that none of the combined significance tests is uniformly most powerful against all alternative hypotheses. Empirical results from simulation (Monte Carlo) studies (e.g., Becker, 1985; George, 1977) provide some guidelines for the selection of a test procedure and show that, in some cases, differences in power among tests are slight. However, optimally one's choice of a statistical procedure should depend on both the nature of the expected (or interesting) outcomes of the series of studies and the behavior of the available tests.

Combined Significance Methods

Rosenthal (1978) reviewed eight combined significance tests, and there are others (e.g., Mudholkar & George, 1979). We present two combined significance methods. One is the method most highly recommended by Rosenthal--the Stouffer method, which is the "method of adding z's" described by Stouffer, Suchman, DeVinney, Star, and Williams (1949). The second method was suggested by Fisher (1932).

We have selected these two tests because power studies (Becker, 1985; Koziol & Perlman, 1978) have shown that these tests perform in a complementary manner. Specifically, the Stouffer test appears to have good power to detect alternatives in which all the populations studied show roughly equal effect sizes. In the validity-study context, this would be a situation in which the increases in validity were roughly the same for all schools or job-groups

studied. Fisher's method has higher power to detect individual (or small numbers of) discrepant populations. Such patterns might arise when the added test batteries increased validity in only a few schools.

Stouffer's Method

Stouffer's test of combined significance (Stouffer et al., 1949) is obtained by summing the standard normal deviates or z values associated with the values p_1 through p_k. The sum is divided by the square root of k (the number of p values), which is the standard deviation of the sum of the k standard normal deviates. The test statistic for this ratio is

$$z_S = \sum_{i=1}^{k} z(p_i)/\sqrt{k}\ , \qquad (4)$$

where $z(p_i) = \Phi^{-1}(p_i)$ represents the standard normal deviate associated with upper-tail probability p_i from the ith study. This test can be computed using the mathematical and statistical functions of programs such as SAS (1990), Minitab (Ryan, Joiner, & Ryan, 1985), or SPSS (1988). FORTRAN programs can also be written to produce the combined significance values. A listing of a SAS program appears in Appendix A.

The statistic in Equation 4 is compared with upper-tail critical values from a table of the standard normal distribution. The test is not conducted as a two-sided test because negative Z_s values do not have a meaningful interpretation in this context. Negative Z_s values result from combinations of negative $z(p_i)$ values, which in turn result from p values larger than 0.5; that is, from nonsignificant individual test results. Thus large negative Z_s values do not represent interesting deviations from the conditions specified by the null hypothesis for the series of studies.

Fisher's Method

A second widely used method for combining probabilities was suggested by Fisher (1932). A related version of this test was also independently described by Pearson (1933). The method requires the transformation of the independent probabilities via the log function. These values are multiplied by the constant -2, which produces (under H_0) a set of identically distributed chi-square variates, each with 2 degrees of freedom. The Fisher test statistic is

$$C_F = -2 \sum_{i=1}^{k} \log (p_i), \qquad (5)$$

which is a chi-square variable with 2k degrees of freedom under H_0. The computation of the Fisher test is also shown in the SAS program in Appendix A. If the probability values associated with the significance test for change in multiple correlation are available, then Fisher's test can also be computed using most spreadsheets (which typically feature the log function).

Use of Multiple Combined Significance Tests

Some authors have recommended the use of several combined significance tests together. Use of multiple combined significance tests and use of combined significance methods together with techniques for pooling or estimating common study results is fairly common, but it leads to slightly elevated levels of Type I error. Elevated error rates occur when several combined significance summaries are applied because they are based on the same data (the p's). Those data are not independent of the estimates of study outcomes (e.g., incremental validities) as well. The usual Bonferroni method (Miller, 1966) can be applied to protect the overall significance level of the set of tests if it is necessary to compute several combined significance summaries.

Validity Studies Based on Independent Samples

Occasionally validity studies compare R or R^2 values computed for independent samples of subjects. In such cases, the test of incremental validity in the individual studies will not be the F-test for change in correlation.

However, combined significance methods can be applied to the probabilities from tests based on independent samples in the same manner described above. The SAS routine in Appendix A would need to be modified by replacing the computed F-test with the X^2 test described for use in individual validity studies based on independent samples. This is the test given in Equation 3. In such cases, the probability values p_1 through p_k would be obtained from the series of X^2 tests from the k schools. Computation of the combined significance tests would proceed exactly as outlined above.

Furthermore, the nonparametric form of the combined significance methods does not preclude combining p's from different validity-study designs (i.e., p's from tests based on dependent samples and p's from independent samples). However, in order for the summaries to be most meaningful, all studies should examine the same hypothesis (or very similar hypotheses) about incremental validity.

POOLING ESTIMATES OF INCREMENTAL VALIDITIES

Correcting Incremental Validities for Artifacts of Restriction of Range and Criterion Unreliability

Although the incremental validity estimates d or d^* may be of interest in a validity study at a single site, they may not be directly comparable across sites. The reason is that these indices of incremental validity in a site are attenuated by range restriction and the unreliability of the criterion variable. Since range restriction and criterion reliability are artifacts of the design of the validity study, one could say that correction of estimates for artifacts is necessary before pooling the estimates, following the validity-generalization tradition (Schmidt & Hunter, 1977).

An alternative characterization is to say that the estimates d and d^* from validity studies at different sites are actually estimating different

quantities. For example, d_1 in study 1 estimates the population value δ_1 of the difference between two (multiple) correlations of test batteries with an unreliable criterion in a restricted population of test scores. The value d_2 in study 2 estimates the population value δ_2 which is the difference between multiple correlations of test scores with a <u>different</u> criterion (and hence criterion reliability) than that of study 1 in a <u>different</u> restricted population than that of study 1. The estimates d_1 from study 1 and d_2 from study 2 estimate conceptually <u>different</u> parameters δ_1 and δ_2. That is, the parameters δ_1 and δ_2 arise as descriptions of different populations.

It does not make sense to pool estimates of different parameters. Hence we would not pool d_1 and d_2 directly. Instead we specify a single parameter that might be estimated from each study. Perhaps the simplest parameter to estimate from each study is the validity increment that would be obtained in the unrestricted population (e.g., the total applicant pool) if the criterion were perfectly reliable. By computing an estimate of this quantity in each study, all studies will be estimating the same conceptual parameter and hence pooling across studies will be sensible. If such a quantity cannot be estimated in each study, an alternative to pooling is the use of nonparametric combined significance summaries, as described above.

Several potential approaches to correction of estimates for unreliability and restriction of range can produce the desired estimates. They involve a combination of the correction for attenuation due to measurement unreliability (e.g., Lord & Novick, 1968, p. 70) with a correction for attenuation due to restriction of range. Perhaps the most elegant correction for the effects of range restriction on correlations is that based on the multivariate correction of the covariance matrix given by Lawley (1943). Unfortunately, it is not easy to derive the effects of this correction on the variance of the "corrected" correlations when the covariances that enter into the correction are themselves uncertain. However, this correction could be used and its effect treated as a multiplicative constant. While this would effectively ignore the uncertainty introduced into the estimated correlations by the correction for range restriction, the effects of this uncertainty are likely to be relatively small. Moreover the Lawley correction permits the operational test scales that are the actual basis of the selection to be treated as explicit selection variables while the criterion and test scales not involved in determining selection are treated as incidental variables whose range is affected by selection on the other variables.

Two other alternatives are less satisfying. One is to estimate range restriction on the criterion (outcome) variable and to correct the correlations via the univariate (Spearman) approach (e.g., Lord & Novick, 1968, p. 145). This approach yields results that are mathematically equivalent to those from the Lawley correction, but it requires knowledge of the criterion variance in both the restricted and unrestricted populations. This is usually unrealistic. A second alternative is to estimate, for each multiple correlation, the linear combination of predictors that has the highest correlation with the criterion (i.e., the predicted score yielded by the regression equation). Compute the variance of this composite in the restricted and in the unrestricted populations, estimate its restriction of

range, and apply the Spearman correction. Since this alternative involves a considerable amount of computation, it is not recommended.

Let $\mathring{R}_1$ and $\mathring{R}_2$ denote the sample multiple correlations R_1 and R_2 after correction for restriction of range via the Lawley correction, and define the relative correction of R_1 and R_2 (which we treat as a known constant for a given site) as

$$c_1 = \mathring{R}_1/R_1$$

and

$$c_2 = \mathring{R}_2/R_2.$$

Let $\hat{P}_1$ and $\hat{P}_2$ be population values of the multiple correlations between test batteries 1 and 2, respectively, and the true score on the criterion in the unrestricted population. Sample estimates $\hat{R}_1$ and $\hat{R}_2$ of $\hat{P}_1$ and $\hat{P}_2$, respectively, are

$$\hat{R}_1 = c_1R_1/\sqrt{\gamma} = \mathring{R}_1/\sqrt{\gamma}, \quad (6)$$

$$\hat{R}_2 = c_2R_2/\sqrt{\gamma} = \mathring{R}_2/\sqrt{\gamma}, \quad (7)$$

where γ is the reliability of the criterion, which is assumed to be known. The corrected correlations $\hat{R}_1$ and $\hat{R}_2$ can be used to construct estimates $\hat{d}$ and $\hat{d}^*$ of incremental validities $\hat{\delta} = \hat{P}_2 - \hat{P}_1$ and $\hat{\delta}^* = \hat{P}_2^2 - \hat{P}_1^2$ in the unrestricted population via

$$\hat{d} = \hat{R}_2 - \hat{R}_1$$

and

$$\hat{d}^* = \hat{R}_2^2 - \hat{R}_1^2.$$

Then

$$\hat{P}_1 = c_1P_1/\sqrt{\gamma},$$

and similarly,

$$\hat{P}_2 = c_2P_2/\sqrt{\gamma}.$$

Correcting Incremental Validities for Shrinkage

Sample estimates of multiple correlations are biased estimators of the population multiple correlation. The bias depends on the sample size n and the number $\underline{a}$ of predictor variables. The bias in R^2 as an estimate of P^2 is approximately

$$\text{BIAS}(R^2) = E(R^2) - P^2 = \frac{a(1-P^2)}{n-1} - \frac{2(n-\underline{a}-1)P^2(1-P^2)}{n^2-1}$$

where the approximation is obtained by ignoring terms proportional to $1/n^2$ (see, e.g., Johnson & Kotz, 1970, p. 244). Because estimates of incremental validity are differences between multiple correlations, we are interested in the bias of estimates of the differences. Given R_1^2 computed with $\underline{a}$ predictors and R_2^2 computed with $\underline{b}$ predictors, and approximating the true squared correlations P_1 and P_2 as $\bar{P}^2 = (P_1^2 + P_2^2)/2$ for the purposes of computing a qualitative estimate of bias,

$$\text{BIAS}(R_2^2 - R_1^2) \approx \frac{(\underline{b}-\underline{a})(1-\bar{P}^2)}{n-1} .$$

If $\bar{P}^2 = .4$, values of $\underline{b}-\underline{a}$ of 4, 5, and 9 imply bias in incremental validity estimates of approximately .0048, .0060, and .0108, respectively, for a study with n = 500, and bias of .0024, .0030, and .0054, respectively, for a study with n = 1000. While these biases are not large in absolute terms, they may not be negligible in terms of the incremental validities of interest. This is particularly true for sample sizes of less than 1000.

If sample sizes of incremental validity studies are less than a few thousand, then a correction for bias is desirable. The correction for shrinkage given by Wherry (1931) or the more complex correction given by Olkin and Pratt (1958) could be applied. Because there is very little difference between the effects of these two corrections, the simpler correction by Wherry may be preferable in practice. Olkin and Pratt note that, because their correction is proportional to 1/n, it has no effect on the large sample distribution of the multiple correlation. This is also true of Wherry's correction. Consequently the large sample variances given here also apply to estimates corrected for shrinkage by either of these two methods.

The Statistical Properties of Incremental Validities

The incremental validity estimates d, d^*, $\hat{d}$, and $\hat{d}^*$ are influenced by sampling variation. Their exact sampling distributions are not known, but large sample approximations have been derived which are quite accurate when the sample sizes are several hundred or larger. It can be shown that in large samples (when the predictor sets are disjoint or the samples are independent), validity increments d, d^*, $\hat{d}$, and $\hat{d}^*$ have normal distributions with means at the true incremental validities (δ, δ^*, $\hat{\delta}$, and $\hat{\delta}^*$, respectively) and variances that can be calculated (estimated) from the matrices of correlations among predictors and criterion.[1] The complexity of the expression for the variance of the incremental validity depends upon whether the two multiple correlations that are used to compute the index are based on different samples and thus can be treated as independent.

---

[1] This holds only if $\delta > 0$ or if the samples are independent or if the predictor sets are disjoint. See Appendix B for an alternative approach when the second predictor set includes the first.

Index d for R_1 and R_2 Computed from Independent Samples

Let n_1 be the size of the sample used to compute R_1 and let n_2 be the size of the sample used to compute R_2. In large samples d has a mean of approximately δ and a variance of approximately

$$\hat{\sigma}^2_{\infty}(d) = \frac{(1 - R_1^2)^2}{n_1} + \frac{(1 - R_2^2)^2}{n_2} .$$

Index d* for R_1 and R_2 Computed from Independent Samples

Let n_1 and n_2 be the sample sizes used to compute R_1 and R_2 respectively. Then in large samples, d^* has a mean of approximately δ^* and a variance of approximately

$$\hat{\sigma}^2_{\infty}(d^*) = \frac{4R_1^2(1 - R_1^2)^2}{n_1} + \frac{4R_2^2(1 - R_2^2)^2}{n_2} .$$

Index d for R_1 and R_2 Computed from the Same Sample

Let n be the sample size. Because R_1 and R_2 are computed from the same sample, they are stochastically dependent and hence

$$\text{Var}(d) = \text{Var}(R_2 - R_1) = \text{Var}(R_2) + \text{Var}(R_1) - 2\text{Cov}(R_1, R_2).$$

Hence in large samples d, has a mean of approximately δ and a variance of approximately

$$\hat{\sigma}^2_{\infty}(d) = \frac{(1 - R_1^2)^2}{n} + \frac{(1 - R_2^2)^2}{n} - \frac{2\text{Cov}_{\infty}(R_1, R_2)}{n} , \qquad (8)$$

The computation of $\text{Cov}_{\infty}(R_1, R_2)$ from the matrix of test and criterion correlations is described starting on page 28 (Result 2).

Index d* for R_1 and R_2 Computed from the Same Sample

Let n be the sample size. Because R_1 and R_2 are computed from the same sample, they are stochastically dependent and hence

$$\text{Var}(d^*) = \text{Var}(R_2^2 - R_1^2) = \text{Var}(R_2^2) + \text{Var}(R_1^2) - 2\text{Cov}(R_1^2, R_2^2).$$

Hence in large samples, d^* has a mean of approximately δ^* and a variance of approximately

$$\hat{\sigma}^2_{\infty}(d^*) = \frac{4R_1^2(1 - R_1^2)^2}{n} + \frac{4R_2^2(1 - R_2^2)^2}{n} - \frac{2\text{Cov}_{\infty}(R_1^2, R_2^2)}{n} , \qquad (9)$$

The computation of $Cov_\infty(R_1^2, R_2^2)$ from the matrix of test and criterion correlations is described starting on page 27 (Result 2).

Index $\hat{d}$ for R_1 and R_2 Computed from Independent Samples

Let n_1 be the size of the sample used to compute R_1 and let n_2 be the size of the sample used to compute R_2. In large samples $\hat{d}$ has a mean of approximately $\hat{\delta}$ and a variance of approximately

$$\hat{\sigma}^2_\infty(\hat{d}) = \frac{c_1^2(1 - R_1^2)^2}{n_1\gamma} + \frac{c_2^2(1 - R_2^2)^2}{n_2\gamma}, \qquad (10)$$

where c_1 and c_2 are correction factors for restriction of range and γ is the reliability of the criterion.

Index $\hat{d}^*$ for R_1 and R_2 Computed from Independent Samples

Let n_1 and n_2 be the sample sizes used to compute R_1 and R_2, respectively. Then in large samples $\hat{d}^*$ has a mean of approximately $\hat{\delta}^*$ and a variance of approximately

$$\hat{\sigma}^2_\infty(\hat{d}^*) = \frac{4c_1^4\, R_1^2(1 - R_1^2)^2}{n_1\gamma^2} + \frac{4c_2^4\, R_2^2(1 - R_2^2)^2}{n_2\gamma^2}, \qquad (11)$$

where c_1 and c_2 are correction factors for restriction of range and γ is the reliability of the criterion.

Index $\hat{d}$ for R_1 and R_2 Computed from the Same Sample

Let n be the sample size. Because R_1 and R_2 are computed from the same sample, they are stochastically dependent and hence

$$Var(\hat{d}) = Var(\hat{R}_2 - \hat{R}_1) = Var(\hat{R}_2) + Var(\hat{R}_1) - 2Cov(\hat{R}_1, \hat{R}_2).$$

Hence in large samples, $\hat{d}$ has a mean of approximately $\hat{\delta}$ and a variance of approximately

$$\hat{\sigma}^2_\infty(\hat{d}) = \frac{c_1^2(1 - R_1^2)^2}{n\gamma} + \frac{c_2^2(1 - R_2^2)^2}{n\gamma} - \frac{2c_1c_2Cov_\infty(R_1, R_2)}{n\gamma}, \qquad (12)$$

where c_1 and c_2 are correction factors for restriction of range and γ is the reliability of the criterion. The computation of $Cov_\infty(R_1, R_2)$ from the matrix of test and criterion correlations is described starting on page 28 (Result 2).

Index $\hat{d}^*$ for R_1 and R_2 Computed from the Same Sample

Let n be the sample size. Because R_1 and R_2 are computed from the same sample, they are stochastically dependent and hence

$$\mathrm{Var}(\hat{d}^*) = \mathrm{Var}(\hat{R}_2^2 - \hat{R}_1^2) = \mathrm{Var}(\hat{R}_2^2) + \mathrm{Var}(\hat{R}_1^2) - 2\mathrm{Cov}(\hat{R}_1^2, \hat{R}_2^2).$$

Hence in large samples, $\hat{d}^*$ has a mean of approximately $\hat{\delta}^*$ and a variance of approximately

$$\hat{\sigma}_\infty^2(\hat{d}^*) = \frac{4\, c_1^4\, R_1^2(1 - R_1^2)^2}{n\gamma^2} + \frac{4\, c_2^4\, R_2^2(1 - R_2^2)^2}{n\gamma^2} - \frac{2c_1c_2\mathrm{Cov}_\infty(R_1^2, R_2^2)}{n\gamma^2}, \quad (13)$$

where c_1 and c_2 are correction factors for restriction of range and γ is the reliability of the criterion. The computation of $\mathrm{Cov}_\infty(R_1^2, R_2^2)$ from the matrix of test and criterion correlations is described starting on page 27 (Result 2).

Combining Estimates of Incremental Validity

Statistical methods for pooling results of incremental validity studies are quite similar regardless of the indexes used to represent incremental validity. All are based on statistical theory for combining asymptotically normal independent estimators (see Hedges, 1983). They are described generically in this section so that they can be applied to either of the indexes ($\hat{d}$ or $\hat{d}^*$) previously discussed.

Suppose that there are k independent validity studies, each of which yields an estimate T of incremental validity with a standard error S(T). Here T may be either of the indexes $\hat{d}$ or $\hat{d}^*$ described previously. Using a subscript to denote the study from which an estimate is obtained, T_i is the estimated incremental validity in the i^{th} study and θ_i is the corresponding incremental validity parameter. Thus the data from k studies is the set of estimates $T_1, \ldots, T_k$ and their standard errors $S(T_1), \ldots, S(T_k)$.

If all of the studies provide estimates of a common incremental validity parameter--that is, if $\theta_1 = \cdots = \theta_k = \theta$--then a weighted linear combination of $T_1, \ldots, T_k$ produces the most precise combined estimate (Hedges, 1983). (See Appendix B for an alternative approach when the second predictor set includes the first.) The optimal linear combination T. involves weighting each T_i by the inverse of its variance $S^2(T_i)$, namely

$$T_\cdot = \frac{\sum_{i=1}^{k} w_i\, T_i}{\sum_{i=1}^{k} w_i}, \quad (14)$$

where $w_i = 1/S^2(T_i)$. When each T_i is based on a large sample, then T. is approximately normally distributed about θ with standard error $\sigma(T_\cdot)$

$$T. \sim N(\theta, \sigma^2(T.)), \tag{15}$$

where

$$\sigma(T.) = [\Sigma^k_{i=1} w_i]^{-1/2} \quad . \tag{16}$$

Thus a test for the statistical significance of the incremental validity uses the test statistic

$$Z = T. / \sigma(T.). \tag{17}$$

If $\theta = 0$, then Z has a standard normal distribution. If Z exceeds the 100α percent critical value of the standard normal distribution, then the incremental validity θ is significantly greater than zero at significance level α. For example, if $Z > 1.64$, the incremental validity is significant at the $\alpha = .05$ level of significance.

A $100(1-\alpha)$ percent confidence interval for the incremental validity θ is given by

$$T. - z_{\alpha/2}\, \sigma(T.) \leq \theta \leq T. + z_{\alpha/2}\, \sigma(T.),$$

where $z_{\alpha/2}$ is the 100α percent two-tailed critical value of the standard normal distribution. For example, if $\alpha = .05$, $z_{\alpha/2} = 1.96$ and a 95 percent confidence interval of θ is given by

$$T. - 1.96\, \sigma(T.) \leq \theta \leq T. + 1.96\, \sigma(T.).$$

Estimating the Variance Across Studies of Incremental Validities

It is convenient to treat the incremental validity parameters as if they were relatively constant across studies; i.e., to assume that $\theta_1 = \cdots = \theta_k$. It may be useful to test this assumption by computing an estimate of the variance (component) of the θ_i's across studies. Formally we may assume that $\theta_1, \ldots, \theta_k$ are a sample from a universe of possible incremental validities. This is consistent with the notion that the particular schools in which validity studies are conducted are a sample from a universe of possible schools, each with its own incremental validity parameter.

A simple estimate of σ^2_θ, the variance of the universe of θ values is

$$\hat{\sigma}^2_\theta = \Sigma^k_{i=1} (T_i - \bar{T})^2 / (k-1) - \Sigma^k_{i=1} s^2(T_i)/k. \tag{18}$$

Note that the first summation is just the usual sample estimate of the variance of $T_1, \ldots, T_k$ and the second term is the average of the variances (squared standard errors) of the T_i. Note also that Equation 18 occasionally yields negative values, which are truncated to zero.

A test of the statistical significance of σ^2_θ (that is, a test of the hypothesis H_0: $\sigma^2_\theta = 0$) uses the statistic

$$H = \sum_{i=1}^{k} \frac{(T_i - T.)^2}{S^2(T_i)}.$$

If $\sigma_\theta^2 = 0$, then H has approximately a chi-squared distribution with (k-1) degrees of freedom. Thus if the computed value of H exceeds the $100(1-\alpha)$ percentage point of the chi-square distribution with k-1 degrees of freedom, σ_θ^2 is significantly greater than zero at significance level α.

It is usually simpler to compute H via the computational formula

$$H = \sum_{i=1}^{k} w_i T_i^2 - \frac{[\sum_{i=1}^{k} w_i T_i]^2}{\sum_{i=1}^{k} w_i}, \qquad (19)$$

where $w_i = 1/S^2(T_i)$. This formula permits computation of H, as well as T. and $\sigma(T.)$, from the sums of the variables w_i, $w_i T_i$, and $w_i T_i^2$.

Power of Pooled Tests for Incremental Validity

The large sample distribution given in Equation 15 can be used along with Equation 17 for the test statistic Z and Equation 16 for $\sigma(T.)$ to obtain the large sample distribution of the parametric test for pooled incremental validity. This yields

$$Z \sim N(\theta/\sigma(T.), 1).$$

Hence the power of the test for incremental validity at significance level α based on the pooled estimate is the probability that a normal random variable with mean $\theta/\sigma(T.)$ and variance 1 exceeds z_α, the 100α percent one-tailed critical value of the standard normal distribution. Thus the power is given by

$$1 - \Phi[z_\alpha - \theta/\sigma(T.)]. \qquad (20)$$

Power computations can be made from Equation 20 whenever the expected validity increment θ is known and the standard errors necessary to compute $\sigma(T.)$ have already been calculated.

Power for R_1 and R_2 Computed from the Same Sample

Let the population validities for the two test batteries be P_1 and P_2, respectively. Let $n_1, \ldots, n_k$ be the total sample sizes in the validity studies. Then, under the assumption stated above, the population value of the incremental validity in each study is $\theta = P_2 - P_1$ and the sampling variance of the estimate $T_i = R_2 - R_1$ in the i^{th} study is

$$S^2(T_i) = \Delta/n_i$$

where $\Delta = (1 - P_1^2)^2 + (1 - P_2^2)^2 - 2\mathrm{Cov}(R_1, R_2)$. The value of Δ is essentially that given in Equation 8, but with known values of P_1 and P_2. Hence the sampling variance $\sigma^2(T.)$ of the pooled estimate of incremental validity is

$$\sigma^2(T.) = \frac{\Delta}{N},$$

where $N = \sum_{i=1}^{k} n_i$ is the total (pooled) sample size across all k studies. This implies that the power of the test for pooled incremental validity is

$$1 - \Phi\left[z_\alpha - \frac{(P_2 - P_1)\sqrt{N}}{\sqrt{\Delta}} \right]. \qquad (21)$$

Note that this estimate of power depends only on the significance level, two validities, Δ, and the total sample size across all k studies. It does not depend directly on the number of predictor variables used to compute R_1 or R_2 but is influenced by them through the covariance of R_1 and R_2 used to compute Δ. Equation 21 can be used to compute power values for a given level of incremental validity whenever Δ can be computed. When the same sample is used to compute R_1 and R_2, the covariance of R_1 and R_2 is not zero. In fact, this covariance is usually quite large, particularly when the incremental validity is small. The reason for this is that the magnitudes of R_1 and R_2 tend to be correlated: If there is little incremental validity, samples that tend to give a large value of R_1 also give a large value of R_2.

However, the magnitude of the correlation between R_1 and R_2 also depends on the difference in the numbers of predictors in models 1 and 2. Specifically, the correlation (and covariance) generally decrease as more variables are included in model 2. For example, a typical value of the correlation between R_1 and R_2 for four added variables is .93 when $P_1 = .40$ and $P_2 = .45$, whereas a typical intercorrelation value for the same population validities would be .91 when the second model includes nine additional variables. Even these seemingly slight differences in correlation values correspond to differences in the power of tests for incremental validity.

The estimate given in Equation 21 of the power of the pooled test for incremental validity was used to compute the power values given in Tables 2 through 4. These computations show that, when the incremental validity is .02 and the total sample size is at least N = 1500, the power of the test exceeds 95 percent when the $\alpha = .05$ level of significance is used and 85 percent when $\alpha = .01$ for nine added variables. When only four variables are added, Table 4 shows that 95 percent power is achieved with less than 500 subjects when $\alpha = .05$, and with less than 750 subjects when $\alpha = .01$.

Because current plans for validity studies include sample sizes substantially larger than the minimum necessary for power of 95 percent for tests at the $\alpha = .05$ level of significance, current studies should have adequate power to detect pooled incremental validities of .02 or even smaller.

Table 2

Power of the Pooled Test for Incremental Validity for Nine Additional Variables as a Function of the Validity Increment and Pooled Sample Size for R_1 and R_2 Computed from the Same Sample and $P_1 = .40$

	Significance Level $\alpha = .05$		Significance Level $\alpha = .01$	
	$P_2 - P_1$		$P_2 - P_1$	
n	.05	.02	.05	.02
250	0.72	0.40	0.45	0.17
500	0.93	0.62	0.79	0.36
750	0.99	0.77	0.93	0.53
1000	1.00	0.87	0.98	0.67
1250	1.00	0.93	1.00	0.78
1500	1.00	0.96	1.00	0.86
1750	1.00	0.98	1.00	0.91
2000	1.00	0.99	1.00	0.94
2200	1.00	0.99	1.00	0.96
2400	1.00	1.00	1.00	0.98
2500	1.00	1.00	1.00	0.98
3000	1.00	1.00	1.00	0.99
3500	1.00	1.00	1.00	1.00
4000	1.00	1.00	1.00	1.00
5000	1.00	1.00	1.00	1.00
6000	1.00	1.00	1.00	1.00
7000	1.00	1.00	1.00	1.00
8000	1.00	1.00	1.00	1.00
9000	1.00	1.00	1.00	1.00
10000	1.00	1.00	1.00	1.00

Note: Power values listed as 1.00 are values greater than .995.

Table 3

Power of the Pooled Test for Incremental Validity for Five Additional Variables as a Function of the Validity Increment and Pooled Sample Size for R_1 and R_2 Computed from the Same Sample and P_1 = .40

	Significance Level α = .05		Significance Level α = .01	
	$P_2 - P_1$		$P_2 - P_1$	
n	.05	.02	.05	.02
250	0.79	0.46	0.55	0.21
500	0.97	0.70	0.87	0.44
750	1.00	0.84	0.97	0.63
1000	1.00	0.92	1.00	0.77
1250	1.00	0.96	1.00	0.86
1500	1.00	0.98	1.00	0.92
1750	1.00	0.99	1.00	0.96
2000	1.00	1.00	1.00	0.98
2200	1.00	1.00	1.00	0.99
2400	1.00	1.00	1.00	0.99
2500	1.00	1.00	1.00	0.99
3000	1.00	1.00	1.00	1.00
3500	1.00	1.00	1.00	1.00
4000	1.00	1.00	1.00	1.00
5000	1.00	1.00	1.00	1.00
6000	1.00	1.00	1.00	1.00
7000	1.00	1.00	1.00	1.00
8000	1.00	1.00	1.00	1.00
9000	1.00	1.00	1.00	1.00
10000	1.00	1.00	1.00	1.00

Note: Power values listed as 1.00 are values greater than .995.

Table 4

Power of the Pooled Test for Incremental Validity for Four Additional Variables as a Function of the Validity Increment and Pooled Sample Size for R_1 and R_2 Computed from the Same Sample and $P_1 = .40$

	Significance Level $\alpha = .05$		Significance Level $\alpha = .01$	
	$P_2 - P_1$		$P_2 - P_1$	
n	.05	.02	.05	.02
250	0.82	0.48	0.59	0.23
500	0.97	0.73	0.90	0.47
750	1.00	0.87	0.98	0.66
1000	1.00	0.94	1.00	0.80
1250	1.00	0.97	1.00	0.89
1500	1.00	0.99	1.00	0.94
1750	1.00	0.99	1.00	0.97
2000	1.00	1.00	1.00	0.98
2200	1.00	1.00	1.00	0.99
2400	1.00	1.00	1.00	1.00
2500	1.00	1.00	1.00	1.00
3000	1.00	1.00	1.00	1.00
3500	1.00	1.00	1.00	1.00
4000	1.00	1.00	1.00	1.00
5000	1.00	1.00	1.00	1.00
6000	1.00	1.00	1.00	1.00
7000	1.00	1.00	1.00	1.00
8000	1.00	1.00	1.00	1.00
9000	1.00	1.00	1.00	1.00
10000	1.00	1.00	1.00	1.00

Note: Power values listed as 1.00 are values greater than .995.

Power for R_1 and R_2 Computed from Independent Samples

Let the population validities for the two test batteries be P_1 and P_2, respectively. Let $n_1, \ldots, n_k$ be the total sample sizes in the validity studies and assume that the two groups within each study are of equal size. The population value of the incremental validity is $\theta = P_2 - P_1$ and the sampling variance of the estimate $T_i = R_2 - R_1$ in the i^{th} study is

$$S^2(T_i) = 2\Delta/n_i$$

where $\Delta = (1 - P_1^2)^2 + (1 - P_2^2)^2$. Note that the covariance term is omitted because R_1 and R_2 are independent. Therefore the numbers of predictors in the two models also do not affect the power of the test for incremental validity when independent samples are used. Thus the sampling variance $\sigma^2(T.)$ of the pooled estimate of incremental validity is

$$\sigma^2(T.) = \frac{2\Delta}{N} \quad ,$$

where $N = \sum_{i=1}^{k} n_i$ is the total (pooled) sample size across all k studies. This implies that the power of the test for pooled incremental validity is

$$1 - \Phi\left[z_\alpha - \frac{(P_2 - P_1)\sqrt{N}}{\sqrt{2\Delta}}\right] \quad . \qquad (22)$$

The estimate given in Equation 22 was used to compute the power values given in Table 5. These computations show that the power of the pooled test for incremental validity is substantially less when the studies use independent samples to compute R_1 and R_2 than when the same sample is used. For example, if the incremental validity is .02 and the $\alpha = .05$ level of significance is used, a total sample size of N = 45,000 is needed to reach a power of 80 percent. If the incremental validity is .01, the power does not attain even 40 percent power for pooled sample sizes as large as N = 50,000.

Table 5

Power of the Pooled Test for Incremental Validity as a Function of the Validity Increment and Pooled Sample Size for R_1 and R_2 Computed from Independent Samples and $P_1 = .40$

	Significance Level $\alpha = .05$		Significance Level $\alpha = .01$	
	P_2-P_1		P_2-P_1	
n	.05	.02	.05	.02
1,000	.25	.10	.09	.03
2,000	.39	.13	.17	.04
3,000	.51	.16	.26	.05
4,000	.61	.19	.34	.06
5,000	.70	.21	.43	.07
6,000	.76	.24	.51	.08
7,000	.82	.26	.59	.09
8,000	.86	.28	.66	.10
9,000	.89	.31	.71	.12
10,000	.92	.33	.76	.13
15,000	.98	.43	.92	.20
20,000	1.00	.52	.98	.26
25,000	1.00	.60	.99	.33
30,000	1.00	.67	1.00	.40
35,000	1.00	.73	1.00	.47
40,000	1.00	.78	1.00	.53
45,000	1.00	.82	1.00	.59
50,000	1.00	.85	1.00	.64

Note: Power values listed as 1.00 are values greater than .995.

THEORETICAL RESULTS

In this section we derive the asymptotic distributions of the incremental validity indexes d and d^*. We use these asymptotic distributions to obtain large sample approximations to the distributions of these indices. We begin by stating a fundamental theorem. Then we use this theorem to obtain the asymptotic joint distribution of the determinants of certain correlation matrices. These distributions are then used to obtain the asymptotic joint distributions of R_1 and R_2, and R_1^2 and R_2^2 which yield the asymptotic distributions of d and d^*.

A Fundamental Theorem

Throughout this section we make use of the multivariate delta method, which follows from the fundamental theorem given below. This theorem is a straightforward generalization of Theorem 4.2.5 in Anderson (1958), as given, for example, in Olkin and Siotani (1976).

Theorem: Let $\mathbf{u}(n) = (u_1(n), \ldots, u_m(n))$ be a vector of random variables such that the limit in probability as $n \to \infty$ of $\mathbf{u}(n) = \mathbf{b} = (b_1, \ldots, b_m)$ and $\sqrt{n}$ $(\mathbf{u}(n) - \mathbf{b})$ is asymptotically normally distributed with mean vector $\mathbf{O}$ and covariance matrix $\mathbf{\Psi}$. If $\mathbf{y}(n) = (y_1(n), \ldots, y_k(n)) = (f_1, \ldots, f_k)$, $k \leq m$, where the $f_i = f_i(\mathbf{u}(n))$ are functions of $\mathbf{u}(n)$ having first and second derivatives in a neighborhood of $\mathbf{u}(n) = \mathbf{b}$, then the asymptotic distribution $\sqrt{n}\,[\mathbf{y}(n) - \mathbf{f}(\mathbf{b})]$ is given by

$$\sqrt{n}\,[\mathbf{y}(n) - \mathbf{f}(\mathbf{b})] \sim N(\mathbf{O}, \mathbf{A}\mathbf{\Psi}\mathbf{A}'),$$

where $\mathbf{A}$ is a k x m matrix with elements

$$a_{ij} = (\partial f_i(\mathbf{u}) / \partial u_j(n)) \Big|_{\mathbf{u} = \mathbf{b}} \quad \text{and} \quad \mathbf{f}(\mathbf{b}) = (f_1(\mathbf{b}), \ldots, f_k(\mathbf{b})).$$

In this paper, the vector $\mathbf{u}(n)$ is composed of correlation coefficients (i.e., $\mathbf{u}$ is the correlation matrix among tests and criteria, arranged as a vector). Because correlation coefficients of multivariate normal variates are functions of sample moments, they have an asymptotic joint distribution that is multivariate normal with a covariance matrix $\mathbf{\Psi}$ which is a function of the population values of the correlations (Pearson & Filon, 1898). Thus the theorem gives a method for computing the asymptotic joint distribution of pairs of multiple correlations, or of any smooth functions of pairs of multiple correlations, such as the indexes of incremental validity considered here.

The Sampling Distribution of Incremental Validities

In this section we apply the fundamental theorem to obtain the sampling distribution of incremental validity indices in large samples. We do so in three steps. First, we obtain the asymptotic joint distribution of the determinants of correlation matrices. We use the joint distribution of the determinants to obtain the asymptotic joint distribution of two multiple correlations and that of two squared multiple correlations. Finally, we use the joint distribution of two multiple correlations to obtain the asymptotic distribution of the incremental validity indices.

Notation

Let X_0, X_1, X_2, ..., X_m be a collection of random variables with a joint multivariate normal distribution, where X_0 represents the criterion variable and X_1, ..., X_m represent predictor variables. We denote the sample and population correlations between X_i and X_j by r_{ij} and ρ_{ij} respectively. We denote a matrix of correlations by defining the set of variables to be correlated. Specifically, for $k \geq 1$, let α_1, ..., α_k denote distinct subsets of the set of integers {0, 1, ..., m}. Then each α_i defines a set of variables--the set of variables whose subscripts are contained in α_i. We use the notation $\mathbf{R}(\alpha_1)$, ..., $\mathbf{R}(\alpha_k)$ to denote the square matrices of correlations of variables implied by the sets α_1, ..., α_k. We will also use the notation $\mathbf{R}(0, \alpha_i)$ to denote the matrix of correlations of X_0 and the variables implied by α_i instead of the more formal $\mathbf{R}(\{0\}, \alpha_i)$.

Result 1: Joint Distribution of Determinants of Correlation Sub-matrices

Let X_0, X_1, ..., X_m be random variables (representing a criterion and test scales) that have a joint multivariate normal distribution. Let α_1, ..., α_k be nonempty sets of the integers between 0 and m inclusive, denoting collections of the m subtests, possibly including the criterion. Thus $\mathbf{R}(\alpha_1)$, ..., $\mathbf{R}(\alpha_k)$ are the sample correlation matrices of the variables implied by α_1, ..., α_k respectively. Then the asymptotic joint distribution of $|\mathbf{R}(\alpha_1)|$, ..., $|\mathbf{R}(\alpha_k)|$, when all of the determinants are computed from correlations based on the same sample of size n, is given by

$$\sqrt{n}\,[(|\mathbf{R}(\alpha_1)|, \ldots, |\mathbf{R}(\alpha_k)|) - (|\mathbf{P}(\alpha_1)|, \ldots, |\mathbf{P}(\alpha_k)|)] \sim \mathbf{N}(\mathbf{O}, \boldsymbol{\Sigma})$$

where $\mathbf{O}$ is a k x 1 vector of zeros and $\boldsymbol{\Sigma}$ is given by (σ_{ij}) and

$$\sigma_{ij} = \sum_{\substack{s \in \alpha_i \\ s<t}} \sum_{t \in \alpha_i} \sum_{\substack{u \in \alpha_j \\ u<v}} \sum_{v \in \alpha_j} 4|\mathbf{P}(\alpha_i)||\mathbf{P}(\alpha_j)|\rho_{(i)}{}^{st}\,\rho_{(j)}{}^{uv} \times$$

$$\left\{ \rho_{st}\,\rho_{uv}\,(\rho_{su}^2 + \rho_{sv}^2 + \rho_{tu}^2 + \rho_{tv}^2)/2 + \rho_{su}\,\rho_{tv} + \rho_{sv}\,\rho_{tu} - (\rho_{st}\,\rho_{su}\,\rho_{sv} + \rho_{st}\,\rho_{tu}\,\rho_{tv} + \rho_{su}\,\rho_{tu}\,\rho_{uv} + \rho_{sv}\,\rho_{tv}\,\rho_{uv}) \right\},$$

where the sums in σ_{ij} are taken so that $s < t$ and $u < v$ and $\rho_{(i)}{}^{st}$ is the element in row s and column t of $\mathbf{P}^{-1}(\alpha_i)$, the <u>inverse</u> of $\mathbf{P}(\alpha_i)$, and $\rho_{ss} = 1$.

A somewhat simpler form of the asymptotic variance (σ_{ii}) of the determinant of a correlation matrix was derived by Olkin and Siotani (1976). We use a slightly more complex expression than needed for the variance terms to define all elements of $\boldsymbol{\Sigma}$ because it generalizes more easily to the situation in which some correlations are known. The covariance terms are obtained by applying the multivariate delta method to the joint distribution of the correlation matrix.

Apply the method by writing the nonredundant elements of the correlation matrix as a vector $r = (r_{01}, r_{02}, \ldots, r_{0m}, r_{12}, \ldots r_{1m}, r_{23}, \ldots, r_{(m-1)m})'$. Then if $|R(\alpha_i)| = f_i(r)$, the asymptotic covariance matrix is computed from the partial derivatives of the f_i with respect to the elements of r. Since $R(\alpha_i)$ is symmetric,

$$\frac{\partial |R(\alpha_i)|}{\partial r_{st}} = 2\, r_{(i)}{}^{st} |R(\alpha_i)|.$$

Hence

$$\sigma_{ij} = 4 \sum_{s<t,\ u<v} |P(\alpha_i)||P(\alpha_j)|\ \rho_{(i)}{}^{st}\ \rho_{(j)}{}^{uv}\ \mathrm{Cov}(r_{st}, r_{uv}), \quad (23)$$

where $\mathrm{Cov}(r_{st}, r_{uv})$ is the covariance of r_{st} and r_{uv} from the asymptotic matrix of r, and $s,t \in \alpha_i$ and $u,v \in \alpha_j$. Substituting the expression given, for example, by Pearson and Filon (1898) for $\mathrm{Cov}(r_{st}, r_{uv})$ yields the result given in the theorem.

Equation 23 also can be written as

$$\sigma_{ij} = \sum_s \sum_t \sum_u \sum_v |P(\alpha_i)||P(\alpha_j)|\ \rho_{(i)}{}^{st}\ \rho_{(j)}{}^{uv}\ \mathrm{Cov}(r_{st}, r_{uv}),$$

where no index is restricted to be less than another.

Using the notation $\rho_{ij,k} = \rho_{ij} - \rho_{ik}\rho_{jk}$,

$$\mathrm{Cov}(r_{st}, r_{uv}) = \tfrac{1}{2}\{ \rho_{su,t}\rho_{tv,u} + \rho_{sv,u}\rho_{tu,s} + \rho_{su,v}\rho_{tv,s} + \rho_{sv,t}\rho_{tu,v} \}.$$

By symmetry,

$$\sigma_{ij} = 2 \sum_s \sum_t \sum_u \sum_v |P(\alpha_i)||P(\alpha_j)|\ \rho_{(i)}{}^{st}\ \rho_{(j)}{}^{uv}\ \rho_{su,t}\ \rho_{tv,u}. \quad (24)$$

When Some Correlations Known. When some of the elements of the $R(\alpha_i)$ are known, the formal computation of the asymptotic distribution remains the same as given above except that each term of the sum involving a known r_{st} vanishes.

When All Correlations Estimated. If all of the correlations are estimated, σ_{ij} can be expressed in the form of simple matrix multiplications, as shown by expanding Equation 24 and carrying out the indicated multiplications by the inverse elements, as follows:

$$\sigma_{ij} = 2|P(\alpha_i)||P(\alpha_j)| \sum_t \sum_u [(\sum_s \rho_{(i)}{}^{st}\ \rho_{su}) - \rho_{tu}][(\sum_v \rho_{(j)}{}^{uv}\ \rho_{tv}) - \rho_{tu}].$$

Let P_{ij} = the correlation matrix between the variables in set α_i and set α_j, with $P_{ii} = P(\alpha_i)$. Let $A = (P_{ii}^{-1} - I)P_{ij}$ and $B = P_{ij}(P_{jj}^{-1} - I)$. Then

$$\sigma_{ij} = 2|P(\alpha_i)||P(\alpha_j)|tr(AB') = 2|P(\alpha_i)||P(\alpha_j)|A \bullet B , \qquad (25)$$

where the dot product operator • denotes the sum of the products of the corresponding elements in the two matrices.

Result 2: Population Covariances of Multiple Correlations Estimated from the Same Sample

Let $X_0, X_1, \ldots, X_m$ be random variables with a joint multivariate normal distribution. Let R_1 be the sample multiple correlation of a subset of variables identified by the set α_1 of indexes with X_0 and let R_2 be the sample multiple correlation of another subset of variables identified by the set α_2 of indexes with X_0, where both correlations are computed from the same sample of size n. Let P_1 and P_2 be the population multiple correlations corresponding to R_1 and R_2 respectively. Then the asymptotic joint distribution of R_1^2 and R_2^2 is given by

$$\sqrt{n}\,[(R_1^2, R_2^2) - (P_1^2, P_2^2)] \sim N(0, \Sigma),$$

where $\Sigma = (\sigma_{ij})$,

$$\sigma_{11} = \frac{c_1}{|P(\alpha_1)|^2} + \frac{c_2|P(0, \alpha_1)|^2}{|P(\alpha_1)|^4} - \frac{2c_3|P(0,\alpha_1)|}{|P(\alpha_1)|^3} ,$$

$$\sigma_{22} = \frac{c_4}{|P(\alpha_2)|^2} + \frac{c_5|P(0, \alpha_2)|^2}{|P(\alpha_2)|^4} - \frac{2c_6|P(0, \alpha_2)|}{|P(\alpha_2)|} ,$$

$$\sigma_{12} = \frac{c}{|P(\alpha_1)||P(\alpha_2)|} ,$$

$$c = \left| c_7 - \frac{c_8|P(0, \alpha_1)|}{|P(\alpha_1)|} - \frac{c_9|P(0, \alpha_2)|}{|P(\alpha_2)|} + \frac{c_{10}|P(0, \alpha_1)||P(0, \alpha_2)|}{|P(\alpha_1)||P(\alpha_2)|} \right| ,$$

$c_1 = \text{Var}\,(|R(0, \alpha_1)|)$, $c_6 = \text{Cov}(|R(\alpha_2)|, |R(0, \alpha_2)|)$,

$c_2 = \text{Var}\,(|R(\alpha_1)|)$, $c_7 = \text{Cov}\,(|R(0, \alpha_1)|, |R(0, \alpha_2)|)$,

$c_3 = \text{Cov}(|R(\alpha_1)|, |R(0, \alpha_1)|)$, $c_8 = \text{Cov}\,(|R(\alpha_1)|, |R(0, \alpha_2)|)$,

$c_4 = \text{Var}\,(|R(0, \alpha_2)|)$, $c_9 = \text{Cov}\,(|R(0, \alpha_1)|, |R(\alpha_2)|)$,

$c_5 = \text{Var}\,(|R(\alpha_2)|)$, $c_{10} = \text{Cov}\,(|R(\alpha_1)|, |R(\alpha_2)|)$,

and the covariances $c_1, \ldots, c_{10}$ are given in Result 1.

The asymptotic joint distribution of R_1 and R_2 is given by

$$\sqrt{n}\,[(R_1, R_2) - (P_1, P_2)] \to N(0, \Psi),$$

where $\Psi = (\Psi_{ij})$,

$$\Psi_{11} = \frac{c_1}{4|P(\alpha_1)|^2 P_1^2} + \frac{c_2|P(0, \alpha_1)|^2}{4|P(\alpha_1)|^4 P_1^2} - \frac{c_3|P(0, \alpha_1)|}{2|P(\alpha_1)|^3 P_1^2},$$

$$\Psi_{22} = \frac{c_3}{4|P(\alpha_2)|^2 P_2^2} + \frac{c_4|P(0, \alpha_2)|^2}{4|P(\alpha_2)|^4 P_2^2} - \frac{c_6|P(0, \alpha_2)|}{2|P(\alpha_2)|^3 P_2^2},$$

$$\Psi_{12} = \frac{C}{4|P(\alpha_1)||P(\alpha_2)|P_1P_2},$$

and $c_1, \ldots, c_6$ and C are given above.

The result is proven by writing the multiple correlations as functions of the determinants of correlation matrices

$$R_1^2 = 1 - \frac{|R(0, \alpha_1)|}{|R(\alpha_1)|},$$

$$R_2^2 = 1 - \frac{|R(0, \alpha_2)|}{|R(\alpha_2)|}.$$

Since R_1 and R_2 are functions of $|R(\alpha_1)|$, $|R(0, \alpha_1)|$, $|R(\alpha_2)|$, and $|R(0, \alpha_2)|$, the joint distributions of (R_1, R_2) and (R_1^2, R_2^2) are derived by applying the delta method using Result 1.

Note that $\Psi_{ij} = \sigma_{ij} / 4P_iP_j$, and hence the correlation

$$\rho(R_1, R_2) = \rho(R_1^2, R_2^2) .$$

We use a more complex expression for the asymptotic variances σ_{ii} and Ψ_{ii} than is strictly necessary because the expression given above generalizes more easily to the case when some of the bivariate correlations are known. If none of the correlations are known then

$$\sigma_{ii} = 4P_i^2 (1 - P_i^2)^2 \text{ and } \Psi_{ii} = (1 - P_i^2)^2.$$

Result 3: Population Variances of Multiple Correlation Differences Estimated from the Same Sample

Let $X_0, X_1, \ldots, X_m$ be random variables with a joint multivariate normal distribution. Let α_1 be a nonempty subset of the set of integers $\{1, 2, \ldots, m\}$ defining (via the subscripts) a subset of the collection of variables. Let α_2 be a distinct nonempty subset of $\{1, 2, \ldots, m\}$ defining another subset of the variables $X_1, \ldots, X_m$. Let R_1 be the sample multiple correlation of X_0 with the variables defined by α_1 and let R_2 be the sample multiple correlation of X_0 with the variables defined by α_2. Let P_1 and P_2 be the population correlations corresponding to R_1 and R_2. Then the asymptotic distribution of $d = R_2 - R_1$ is given by

$$\sqrt{n}\,(d - \delta) \sim N(0, \sigma_d^2) ,$$

where $\delta = P_2 - P_1$,

$$\sigma_d^2 = \Psi_{11} + \Psi_{22} - 2\Psi_{12}$$

and the Ψ_{ij} are given in Result 2. The asymptotic distribution of $d^* = R_2^2 - R_1^2$ is given by

$$\sqrt{n}\,(d^* - \delta^*) \sim N(0, \sigma_{d*}^2),$$

where $\delta^* = P_2^2 - P_1^2$,

$$\sigma_{d*}^2 = \sigma_{11} + \sigma_{22} - 2\sigma_{12} \qquad (26)$$

and the σ_{ij} are given in Result 2.

This result is obtained directly by applying the delta method using the asymptotic distribution given in Result 2.

Result 4: Population Variances of Multiple Correlation Differences Estimated from Independent Samples

Let $X_0, X_1, \ldots, X_m$ be random variables with a joint multivariate normal distribution and let α_1 and α_2 be distinct subsets of the integers $\{1, \ldots, m\}$ such that each defines a distinct set of the variables $X_1, \ldots, X_m$. Let R_1 be the multiple correlation of X_0 with the variables defined in α_1 computed from a sample of size n_1 and let R_2 be the multiple correlation of X_0 with the variables defined by α_2 computed from an <u>independent sample</u> of size n_2. Let P_1 and P_2 be the population multiple correlations corresponding to R_1 and R_2. Then if $n = n_1 + n_2$ and $\pi_1 = n_1/n$ and $\pi_2 = n_2/n$ remain fixed as $n \to \infty$, the asymptotic distributions of $d = R_2 - R_1$ and $d^* = R_2^2 - R_1^2$ are given by

$$\sqrt{n}\,(d - \delta) \sim N(0, \sigma_d^2)$$

and

$$\sqrt{n}\ (d^* - \delta^*) \ \sim\ N(0,\ \sigma^2_{d*})$$

where $\delta = P_2 - P_1$, $\delta^* = P_2^2 - P_1^2$, and

$$\sigma^2_d = \frac{(1 - P_1^2)^2}{\pi_1} + \frac{(1 - P_2^2)^2}{\pi_2},$$

$$\sigma^2_{d*} = \frac{4P_1^2(1 - P_1^2)^2}{\pi_1} + \frac{4P_2^2(1 - P_2^2)^2}{\pi_2}.$$

This result follows directly from the asymptotic distributions of $\sqrt{n}_i(R_i - P_i)$ and $\sqrt{n}_i(R_i^2 - P_i^2)$ and the statistical independence of R_1 and R_2.

Using the Theoretical Results with Estimated Variances

Results 3 and 4 give the asymptotic distributions of incremental validities in which the asymptotic variance is a function of the matrix of population correlations among variables. Thus this distribution theory is of little use when (as in any real application) the entire matrix of population correlations is not known. To use these results, it is necessary to show that estimating the asymptotic variances from sample correlations still yields a valid asymptotic distribution.

Result 5: Sample Variances of Multiple Correlation Differences Estimated from the Same Sample

Suppose that the conditions stated in Result 3 obtain. Define $\hat{\sigma}^2_d$ and $\hat{\sigma}^2_{d*}$ as the estimates of σ^2_d and σ^2_{d*} that would be computed by using the corresponding sample correlation coefficients in place of the population correlations. Hence $\hat{\sigma}^2_d$ and $\hat{\sigma}^2_{d*}$ are random variables depending on the sample correlations. Then the following asymptotic distributions hold as $n \to \infty$:

$$\sqrt{n}\ (d - \delta)\ /\ \hat{\sigma}_d \ \sim\ N(0,\ 1),$$

and

$$\sqrt{n}\ (d^* - \delta^*)\ /\ \hat{\sigma}_{d*} \ \sim\ N(0,1).$$

These asymptotic distributions imply that, in large samples,

$$d \sim N(\delta,\ \hat{\sigma}^2_d/\ n)$$

and

$$d^* \sim N(\delta^*,\ \hat{\sigma}^2_{d*}/\ n).$$

Result 6: Sample Variances of Multiple Correlation Differences Estimated from Independent Samples

Suppose that the conditions of Result 4 obtain. Define $\hat{\sigma}_d^2$ and $\hat{\sigma}_{d*}^2$ as the estimates of σ_d^2 and σ_{d*}^2 obtained by substituting the stochastically independent sample multiple correlations R_1 and R_2 for the population multiple correlations P_1 and P_2 respectively. Then the following asymptotic distributions hold as $n \to \infty$

$$\sqrt{n}\ (d - \delta) \ / \ \hat{\sigma}_d \ \sim \ N(0, 1)$$

and

$$\sqrt{n}\ (d^* - \delta^*) \ / \ \hat{\sigma}_{d*} \ \sim \ N(0, 1).$$

These asymptotic distributions imply that, in large samples,

$$d \ \sim \ N(\delta, \ \hat{\sigma}_d^2 \ / \ n)$$

and

$$d^* \ \sim \ N(\delta^*, \ \hat{\sigma}_{d*}^2 \ / \ n).$$

Results 5 and 6 follow from Results 3 and 4 by noting that the sample correlation matrix **R** converges in probability to the population correlation matrix **P** and σ_d and σ_{d*} are all continuous functions of the elements of **P** (see, e.g., Rao, 1973, p. 385, Theorem 6a.2(i)).

Results When Some Correlations Are Known

In some situations, some of the correlations will be known with a very high degree of precision. For example, if a test battery has been widely used for some extended period, the correlations among tests in the battery may be essentially known. That is, for some r_{ij}, we may know the value of the corresponding population correlation ρ_{ij}. In such cases, it is desirable to increase the precision of estimates of incremental validity by utilizing the fact that some of the correlations are known.

We compute estimates of multiple correlations and incremental validity when some correlations are known by substituting the values of the known correlations for their sample estimates. This procedure yields consistent estimates of the multiple correlations under the model with some known correlations, but the estimates so derived are not the maximum likelihood estimates (see Olkin & Sylvan, 1977). One explanation is that the maximum likelihood estimates (MLEs) of joint covariance matrices are rather complex when some correlations are known, which in turn yield rather complicated (or intractable) expressions for the MLEs of the multiple correlations. The strategy suggested here has the advantages that it produces consistent estimates with reduced variance when some correlations are known (compared to the situation when all correlations must be estimated), it is quite flexible as to patterns of known correlations that can be handled, and it can be further generalized to cases where data from an independent sample are pooled together to strengthen estimates.

Results 1, 2, 3, and 4 generalize directly to the case where the correlations are known. In the case where all correlations were estimated, we derived the asymptotic distributions of functions (e.g., determinants and multiple correlations) from those estimated correlations. When some correlations are known we consider functions of both the estimated correlations and the known

correlations. The key to the generalization of results is the recognition that, since a known correlation ρ_{ij} is a fixed constant, its variance and covariance with any other quantity must be zero. Also any function of all fixed arguments must also be a fixed constant. Using this idea, the generalization of Result 1 is given below as Result 7.

Result 7: Generalization of Result 1 where Some Correlations Known

Let $X_0, X_1, \ldots, X_m$ be random variables (representing criterion and test scales) that have a joint multivariate normal distribution. Let $\alpha_1, \ldots, \alpha_k$ be nonempty sets of 0 through m inclusive, denoting collections of the m subtests. Let $R(\alpha_1), \ldots, R(\alpha_k)$ be the correlation matrices of the variables implied by $\alpha_1, \ldots, \alpha_k$ respectively, where at least one of the elements of each $R(\alpha_i)$ is a sample correlation and the others are population correlations. For each i=1, ..., k define a status-indicator matrix $K(\alpha_i)$ with the same dimensions as $R(\alpha_i)$, but where the elements $k_{(i)st}$ of $K(\alpha_i)$ are defined as 0 or 1 depending upon whether the corresponding element $r_{(i)st}$ of $R(\alpha_i)$ is known. Specifically,

$$k_{(i)st} = \begin{cases} 0 & \text{if } r_{(i)st} = \rho_{(i)st} \text{ is known} \\ 1 & \text{if } r_{(i)st} \text{ is estimated.} \end{cases}$$

for $s, t \in \alpha_i$. Then the asymptotic joint distribution of $|R(\alpha_1)|, \ldots, |R(\alpha_k)|$ when all of the determinants are computed from correlations based on the same sample of size n is given by

$$\sqrt{n}\,[(|R(\alpha_1)|, \ldots, |R(\alpha_k)|) - (|P(\alpha_1)|, \ldots, |P(\alpha_k)|)] \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ is given by (σ_{ij}) and

$$\sigma_{ij} = \sum_{\substack{s\in\alpha_i \\ s<t}} \sum_{t\in\alpha_i} \sum_{\substack{u\in\alpha_j \\ u<v}} \sum_{v\in\alpha_j} 4|P(\alpha_i)||P(\alpha_j)| k_{(i)st}\, k_{(j)uv}\, \rho_{(i)}{}^{st}\, \rho_{(j)}{}^{uv} \times$$

$$\left\{ \rho_{st}\,\rho_{uv}\,(\rho_{su}^2 + \rho_{sv}^2 + \rho_{tu}^2 + \rho_{tv}^2)/2 + \rho_{su}\,\rho_{tv} + \rho_{sv}\,\rho_{tu} - (\rho_{st}\,\rho_{su}\,\rho_{sv} + \rho_{st}\,\rho_{tu}\,\rho_{tv} + \rho_{su}\,\rho_{tu}\,\rho_{uv} + \rho_{sv}\,\rho_{tv}\,\rho_{uv}) \right\}$$

where the sums in σ_{ij} are taken so that $s < t$ and $u < v$, and $\rho_{(i)}{}^{st}$ is the element in row s and column t of $P^{-1}(\alpha_i)$, the inverse of $P(\alpha_i)$, and $\rho_{ss} = 1$.

Result 8: Generalization of Result 2 where Some Correlations Known

Let $X_0, X_1, \ldots, X_m$ be random variables with a joint nonsingular multivariate normal distribution. Let R_1 be the sample estimate of the multiple correlation with X_0 of a subset of variables identified by the set of indices α_1 and let R_2 be the sample estimate of the multiple correlation with X_0 of a subset

of variables identified by the set of indices α_2. Some (but not all) of the bivariate correlations may be known. Thus the corresponding population correlations may be substituted for the corresponding sample correlations in the computation of R_1 and R_2. Whenever sample correlations are used to compute R_1 or R_2, they are based on the same sample of size n. Let the status indicator functions $L(\alpha_i)$ and $L(0, \alpha_i)$ be defined so that

$$L(\alpha_i) = \begin{cases} 0 & \text{if all elements of } \mathbf{R}(\alpha_i) \text{ are known} \\ 1 & \text{if at least one element of } \mathbf{R}(\alpha_i) \text{ is estimated} \end{cases}$$

and

$$L(0, \alpha_i) = \begin{cases} 0 & \text{if all elements of } \mathbf{R}(0, \alpha_i) \text{ are known} \\ 1 & \text{if at least one element of } \mathbf{R}(0, \alpha_i) \text{ is estimated.} \end{cases}$$

Then the asymptotic joint distribution of R_1^2 and R_2^2 is given by

$$\sqrt{n}[(R_1^2, R_2^2) - (P_1^2, P_2^2)] \sim N(0, \Sigma)$$

where $\Sigma = (\sigma_{ij})$ and

$$\sigma_{11} = \frac{c_1 L(0, \alpha_1)}{|\mathbf{P}(\alpha_1)|^2} + \frac{c_2 L(\alpha_1)|\mathbf{P}(0, \alpha_1)|^2}{|\mathbf{P}(\alpha_1)|^4} - \frac{2c_3 L(\alpha_1)L(0, \alpha_1)|\mathbf{P}(0, \alpha_1)|}{|\mathbf{P}(\alpha_1)|^3},$$

$$\sigma_{22} = \frac{c_3 L(0, \alpha_2)}{|\mathbf{P}(\alpha_2)|^2} + \frac{c_4 L(\alpha_2)|\mathbf{P}(0, \alpha_2)|^2}{|\mathbf{P}(\alpha_2)|^4} - \frac{2c_6 L(\alpha_2)L(0, \alpha_2)|\mathbf{P}(0, \alpha_2)|}{|\mathbf{P}(\alpha_2)|^3},$$

$$\sigma_{12} = \frac{C}{|\mathbf{P}(\alpha_1)||\mathbf{P}(\alpha_2)|},$$

$$C = \Bigg| c_7 L(0, \alpha_1) L(0, \alpha_2) + \frac{c_8 L(\alpha_1) L(0, \alpha_2)|\mathbf{P}(0, \alpha_1)|}{|\mathbf{P}(\alpha_1)|}$$

$$+ c_9 \frac{L(0, \alpha_1) L(\alpha_2)|\mathbf{P}(0, \alpha_2)|}{|\mathbf{P}(\alpha_2)|} + c_{10} \frac{L(\alpha_1) L(\alpha_2)|\mathbf{P}(0, \alpha_1)||\mathbf{P}(0, \alpha_2)|}{|\mathbf{P}(\alpha_1)||\mathbf{P}(\alpha_2)|} \Bigg|,$$

$c_1 = \mathrm{Var}(|R(0, \alpha_1)|)$, $c_6 = \mathrm{Cov}(|R(\alpha_2)|, |R(0, \alpha_2)|)$,

$c_2 = \mathrm{Var}(|R(\alpha_1)|)$, $c_7 = \mathrm{Cov}(|R(0, \alpha_1)|, |R(0, \alpha_2)|)$,

$c_3 = \mathrm{Cov}(|R(\alpha_1)|, |R(0, \alpha_1)|)$, $c_8 = \mathrm{Cov}(|R(\alpha_1)|, |R(0, \alpha_2)|)$,

$c_4 = \mathrm{Var}(|R(0, \alpha_2)|)$, $c_9 = \mathrm{Cov}(|R(0, \alpha_1)|, |R(\alpha_2)|)$,

$c_5 = \mathrm{Var}(|R(\alpha_2)|)$, $c_{10} = \mathrm{Cov}(|R(\alpha_1)|, |R(\alpha_2)|)$,

and the covariances $c_1, \ldots, c_{10}$ are given in Result 7. The asymptotic distribution of R_1 and R_2 is given by

$$\sqrt{n}[(R_1, R_2) - (P_1, P_2)] \sim N(0, \Psi),$$

where $\Psi = (\Psi_{ij})$,

$$\Psi_{11} = \frac{c_1 L(0, \alpha_1)}{4|P(\alpha_1)|^2 P_1^2} + \frac{c_2 L(\alpha_1)|P(0, \alpha_1)|^2}{4|P(\alpha_1)|^4 P_1^2} - \frac{c_3 L(\alpha_1) L(0, \alpha_1)|P(0, \alpha_1)|}{2|P(\alpha_1)|^3 P_1^2},$$

$$\Psi_{22} = \frac{c_3 L(0, \alpha_2)}{4|P(\alpha_2)|^2 P_2^2} + \frac{c_4 L(\alpha_2)|P(0, \alpha_2)|^2}{4|P(\alpha_2)|^4 P_2^2} - \frac{c_6 L(\alpha_2) L(0, \alpha_2)|P(0, \alpha_2)|}{|P(\alpha_2)|^3 P_2^2},$$

$$\Psi_{12} = \frac{C}{4|P(\alpha_1)||P(\alpha_2)|P_1 P_2},$$

and $c_1, \ldots, c_6$ and C are given above.

Results 3, 4, 5, and 6 are correct as stated for the case of some known correlations, provided that the covariance matrix for the joint distribution of (R_1, R_2) derived via Results 7 and 8 is used in place of that given in Results 1 and 2.

Note

Although Results 7 and 8 provide a method to increase precision of estimates by using known values of intercorrelations among predictor variables, extensive computations have shown that it produces only a small increase. The use of the method given in these results is computationally rather involved and could thus be justified only if sample sizes were quite marginal. If the overall power of tests for pooled incremental validity is adequate, the additional

precision afforded by the use of these methods does not justify the added computational complexity.

SUMMARY OF PROCEDURES FOR SYNTHESIZING INCREMENTAL VALIDITY RESULTS

This section is a practical guide to procedures for synthesizing the results of incremental validity studies. It provides a step-by-step listing of procedures to be followed for both estimation of incremental validity across studies and testing of the combined significance of the results. An example based on hypothetical results from four schools demonstrates the application of the procedures[2].

Step I: Conduct the Incremental Validity Study at Each Site

At each site (school) the incremental validity study compares a sample validity R_1 with another sample validity R_2 to determine whether R_2 is larger than R_1. Formally this involves a test of the hypothesis that the population validity P_2 associated with R_2 exceeds the population validity P_1 associated with R_1; that is, a test of the hypothesis

$$H_0 : P_2 = P_1.$$

or the identical test that $P_2^2 = P_1^2$. The details of the hypothesis test depend on whether dependent or independent samples are used to compute R_1 and R_2, as discussed above.

R_1 and R_2 Computed from the Same Sample

Case 1. If R_1 and R_2 are computed from the same sample and the predictors for R_1 are a subset of the predictors for R_2, then the appropriate test for incremental validity is the usual F-test for change in multiple correlation. Let $\underline{a}$ be the number of tests used as predictors in R_1 and let $\underline{b} > \underline{a}$ be the number of tests used as predictors in R_2, and let n be the sample size. Compute the F-test given in Equation 1 and compare it to the critical value for an F-distribution with $(\underline{b} - \underline{a})$ and $(n - \underline{b} - 1)$ degrees of freedom. Reject the hypothesis of no incremental validity if the computed value of F exceeds the critical value.

Case 2. If R_1 and R_2 are computed from the same sample but one set of predictors is not a subset of the other, the usual F-test for change in multiple correlation cannot be used. Compute the test statistic X^2 given in Equation 2. Reject the hypothesis of no incremental validity at significance level α if the

---

[2] The methods described in Steps IV through VII are less accurate than the methods of Appendix B when the second predictor set includes the first, as it does in the examples. The methods described in these sections are valid if the predictor sets are disjoint or if the samples are independent.

computed value of X^2 exceeds the 100(1-α) percentile point of the chi-squared distribution with one degree of freedom.

R_1 and R_2 Computed from Independent Samples

If R_1 and R_2 are computed from independent samples the usual F-test for change in multiple correlation cannot be used. Compute the test statistic X^2 given in Equation 3. Reject the hypothesis of no incremental validity at significance level α if the computed value of X^2 exceeds the 100(1-α) percentile point of the chi-squared distribution with one degree of freedom.

Example

Table 6 shows a small data set representing the results of validity studies on four independent samples or schools. Separate regressions (using batteries 1 and 2) have been conducted for each school to obtain the values of R_1 and R_2 for each single sample of subjects. Table 6 shows the differences in squared correlations that lead to the individual significance (F) tests. For this example we have used a = 10 and b = 20 for all four schools or studies. (Either a or b or both could vary across studies, however.)

Each school's F-test is presented in Table 7, with upper-tail p values (in the second column) indicating that significant increases in validity are found for two of the four schools. The probabilities ranged from .004 to .539. Two of the results are "significant" by traditional standards (i.e., $\alpha < .05$).

Table 6
Example: Data

School	n	R_1 (a)	R_2 (b)	$R_2^2 - R_1^2$
Air Traffic Controller	470	.400 (10)	.420 (20)	.016
Fire Control Technician	530	.380 (10)	.424 (20)	.036
Gunner's Mate	700	.440 (10)	.473 (20)	.030
Electrician's Mate	460	.250 (10)	.290 (20)	.022

Step II: Compute Tests of Combined Significance of Incremental Validity

The validity study conducted at each site will have provided a significance test as described in Step I. From each study's significance test, an upper-tailed probability is obtained. These values p_1 through p_k should then be used to compute either Stouffer's (Stouffer et al., 1949) or Fisher's (1932) combined significance test, depending on the expected outcomes of interest.

Stouffer's test, given in Equation 4, may be somewhat more likely to detect the outcome in which all sites show roughly equal-sized increments to validity. Fisher's test (Equation 5) should be used if the question of added validity for any population is of interest. The hypothesis of no increment to validity in any

population is rejected at level α if the selected test exceeds the $100(1-\alpha)$ percent critical value in the appropriate reference distribution.

Example

Table 7 shows the values of the transformed p's used in the two combined significance tests. The values for the normal deviates ($z(p_i)$) and the log-transformed p's were obtained using the mathematical and probability functions of the Minitab mainframe-computer package (Ryan et al., 1985).

Table 7
Example: Computation of Significance Tests

School	F	(df)	p_i	$z(p_i)$	$\log(p_i)$
Air Traffic Controller	0.894	(10, 449)	.539	-0.097	-0.62
Fire Control Technician	2.220	(10, 509)	.016	2.155	-4.16
Gunner's Mate	2.600	(10, 679)	.004	2.636	-5.47
Electrician's Mate	1.059	(10, 439)	.393	0.272	-0.93
Totals				4.965	-11.19

The Stouffer value, which equals 2.48, is significant compared to the standard normal distribution ($p = .007$). The Fisher value of 22.37 is compared to the chi-square distribution with $2k = 8$ degrees of freedom. The observed level of significance for the Fisher test was .0043, only slightly smaller than the probability for the Stouffer test. Both are significant at even the relatively stringent $\alpha = .01$ significance level.

Both tests indicate that the null model, of no increment to validity in any population studied, should be rejected. The additional test battery <u>does</u> add to validity in at least one of the populations studied. The combined significance methods cannot identify <u>which</u> population or populations show this added validity, however.

Step III: Obtain Information for Artifact Correction in Each Study

In order to correct the incremental validity in a study for the artifacts of unreliability and restriction of range, two pieces of information are needed. One is the criterion reliability. The other is the ratio u of the standard deviation of the test score in the unrestricted population to the standard deviation in the study. The unrestricted population must, of course, be defined in the same way for all studies. A good choice for the unrestricted population would be the general applicant pool that takes the ASVAB. No example of artifact correction is provided here.

Step IV: Compute the Index of Incremental Validity and its Variance for Each Study

In order to combine the incremental validities across studies, it is necessary to compute the index of incremental validity and its sampling variance

in each study. The entire process should be done once for the index $\hat{d}$ of change in multiple correlations and once for the index $\hat{d}^*$ of change in squared multiple correlations. First compute the indexes of artifact-corrected incremental validity

$$\hat{d} = \hat{R}_2 - \hat{R}_1$$

and

$$\hat{d}^* = \hat{R}_2^2 - \hat{R}_1^2 \quad ,$$

using the formulas given in Equations 6 and 7. The sampling variances of these indexes depend on whether R_1 and R_2 are computed from the same sample or from independent samples.

If R_1 and R_2 are computed from the same sample in a particular site, use the formulas given in Equations 12 and 13 to compute the sampling variances of $\hat{d}$ and $\hat{d}^*$. If R_1 and R_2 are computed from independent samples, use the formulas given in Equations 10 and 11 to compute the sample variance of $\hat{d}$ and $\hat{d}^*$.

If artifact corrections are not used, then c_1, c_2, and γ are all assigned a value of 1 in Equations 10, 11, 12, or 13 when computing the sampling variance incremental validity.

Example

Table 8 shows the multiple correlations R_1 and R_2, the covariances between R_1 and R_2, and the estimates of d and their variances for the four hypothetical schools. Analogous values for d^* (the difference in squared correlations), and covariances between R_1^2 and R_2^2 are shown in Table 9. Because the data are from one sample within each school and artifact corrections were not applied, Equations 12 and 13 were used to compute the variances with the values of c_1, c_2, and γ set to 1.0.

Table 8
Example: Estimates and Variances of Differences in Correlations

School	R_1	R_2	$d = R_2 - R_1$	$Cov(R_1, R_2)$	$\hat{\sigma}^2(d)$	$\hat{\sigma}(d)$
Air Traffic Controller	0.400	0.420	0.020	0.66	0.0001	0.0117
Fire Control Technician	0.380	0.424	0.044	0.63	0.0003	0.0166
Gunner's Mate	0.440	0.473	0.033	0.56	0.0002	0.0137
Electrician's Mate	0.250	0.290	0.040	0.76	0.0004	0.0208

Table 9
Example: Estimates and Variances of Differences in Squared Correlations

School	R_1^2	R_2^2	$d^* = R_2^2 - R_1^2$	$Cov(R_1^2, R_2^2)$	$\hat{\sigma}^2(d^*)$	$\sigma(d^*)$
Air Traffic Controller	0.160	0.176	0.016	0.445	0.0001	0.0092
Fire Control Technician	0.144	0.180	0.036	0.417	0.0001	0.0116
Gunner's Mate	0.194	0.224	0.030	0.464	0.0002	0.0128
Electrician's Mate	0.062	0.084	0.022	0.220	0.0001	0.0115

It is also possible to compute confidence intervals using the d and d^* estimates from the four studies. Table 10 shows 95 percent confidence intervals for the population differences in correlations ($P_2 - P_1$) and for the squared differences ($P_2^2 - P_1^2$) for the four schools in the example. These confidence intervals also provide an alternative method of testing the null hypothesis of no incremental validity in each study. However, for smaller samples, the usual F-test will be more accurate since the confidence intervals are based on asymptotic (large-sample) results.

Note that negative values of $P_2 - P_1$ are impossible when battery 2 includes battery 1 (and they may be highly implausible in other circumstances). If negative lower confidence limits are computed in such circumstances, they should be truncated to zero.

Table 10
Example: Ninety-five Percent Confidence Intervals for Incremental Validities

	$\delta = P_2 - P_1$		$\delta^* = P_2^2 - P_1^2$	
School	Lower limit	Upper limit	Lower limit	Upper limit
Air Traffic Controller	-0.003	0.043	-0.002	0.035
Fire Control Technician	0.012	0.076	0.013	0.059
Gunner's Mate	0.006	0.060	0.005	0.055
Electrician's Mate	-0.001	0.081	-0.000	0.045

Note: Negative values are included to illustrate computations.

Step V: Calculate the Variance Across Studies of the Population Values of the Incremental Validities in the Unrestricted Population

Compute the estimate of the variance across studies of the population values of the incremental validities in the unrestricted populations using the formula given in Equation 18. Carry out the analysis once for $\hat{d}$ and once for $\hat{d}^*$. That is, if $\hat{d}_1, \ldots, \hat{d}_k$ are the $\hat{d}$ indices from the k studies to be combined, let

$$T_1 = \hat{d}_1, \; T_2 = \hat{d}_2, \; \ldots, \; T_k = \hat{d}_k$$

and

$$S^2(T_1) = \hat{\sigma}^2_{\infty}(\hat{d}_1), \; S^2(T_2) = \hat{\sigma}^2_{\infty}(\hat{d}_2), \; \ldots, \; S^2(T_k) = \hat{\sigma}^2_{\infty}(\hat{d}_k)$$

and apply the formula given in Equation 18. Then carry out the same process with the $\hat{d}^*$ values. If $\hat{d}_1^*, \ldots, \hat{d}_k^*$ are the $\hat{d}^*$ indices from the k studies, use Equation 18 with

$$T_1 = \hat{d}_1^*, \; \ldots, \; T_k = \hat{d}_k^*$$

and

$$S^2(T_1) = \hat{\sigma}^2_{\infty}(\hat{d}_1^*), \; \ldots, \; S^2(T_k) = \hat{\sigma}^2_{\infty}(\hat{d}_k^*).$$

To test the hypothesis that the incremental validity varies across studies, compute the test statistic H using the computational formula given in Equation 19.

Example

Again our example is based on uncorrected estimates d and d^*. For both the d and d^* estimates, the variance components computed using Equation 18 were estimated to be zero. This suggests that there is no variation in the parameters representing incremental validity, when either differences or squared differences in multiple correlations are used. All populations under study can be considered to show the same increment in validity due to the added predictor variables.

In fact, both $\hat{\sigma}^2_\theta$ values were actually slightly negative, though they were very small. Actual values were -0.00015 for d and -0.00006 for d^*. Conventionally, however, such negative variance-component estimates are truncated to zero.

Table 11 shows the terms used in the computation of the H statistics (also the pooled estimates and their standard errors) for the two incremental validity measures. The weight terms for the two measures (labeled w_i and w_i^*) are computed as the inverses of the variances of each school's estimates. Each weight is then multiplied by its respective incremental validity estimate and the square of the estimate, as shown in Table 11.

Table 11
Example: Computation of the Summary Statistics for Two Incremental Validity Indices

	$d = R_2 - R_1$			$d^* = R_2^2 - R_1^2$		
School	Weight(w_i)	$w_i d_i$	$w_i d_i^2$	Weight(w_i^*)	$w_i^* d_i^*$	$w_i^* d_i^{*2}$
Air Traffic Controller	7353.29	147.07	2.94	11690.25	191.72	3.14
Fire Control Technician	3642.53	160.27	7.05	7376.92	263.92	9.44
Gunner's Mate	5293.70	174.69	5.76	6066.33	180.35	5.36
Electrician's Mate	2314.85	92.59	3.70	7616.13	168.32	3.72
Totals	18604.37	574.62	19.46	32749.64	804.30	21.67

The homogeneity test statistics for both measures of incremental validity also support the finding of consistency in the magnitudes of the population parameters. In each case, the test statistic H has a chi-square distribution with three degrees of freedom under the null hypothesis of no variation in population parameters. Using Equation 19, the homogeneity test is H = 1.714 (p = .63, df = 3) for the differences in multiple correlations (i.e., the value computed using the d estimates). The value of the test for the squared multiple correlations is H = 1.915 (p = .59, df = 3). Neither value is significant even at the most lenient conventional significance level (e.g., α = .10).

Step VI: Calculate the Combined Estimate of Incremental Validity

The combined estimate of incremental validity is a weighted average of the values from the individual studies. Compute the combined (weighted average) estimate across studies of $\hat{d}$ and $\hat{d}^*$ separately using the formula given in

Equation 14, and use the formula in Equation 16 to compute the standard error of each weighted average. First, let

$$T_1 = \hat{d}_1, \ldots, T_k = \hat{d}_k$$

and

$$s^2(T_1) = \hat{\sigma}^2_{\infty}(\hat{d}_1), \ldots, s^2(T_k) = \hat{\sigma}^2_{\infty}(\hat{d}_k)$$

to combine the $\hat{d}$ values. Then let

$$T_1 = \hat{d}^*_1, \ldots, T_k = \hat{d}^*_k$$

and

$$s^2(T_1) = \hat{\sigma}^2_{\infty}(\hat{d}^*_1), \ldots, s^2(T_k) = \hat{\sigma}^2_{\infty}(\hat{d}^*_k)$$

to combine the $\hat{d}_*$ values.

Example

The combined estimates of incremental validity are weighted averages of the differences and of the squared differences shown in Tables 6 and 7. For both measures of added validity, these averages can be considered to represent common parameters, since the null hypothesis of no variation was retained for both. The average weighted difference in multiple correlations is 0.031 with a standard error of 0.007. The average difference in squared multiple correlations is 0.025 with a standard error of 0.006.

Step VII: Compute a Confidence Interval for the Incremental Validity

Use the pooled estimate of incremental validity and its standard error computed in Step VI to compute a confidence interval for the incremental validity. If this confidence interval does not contain zero or, equivalently, if the test given in Equation 17 leads to a significant Z value, reject the hypothesis of zero incremental validity.

Example

Ninety-five percent confidence intervals were computed using the two pooled estimates of the validity increment. For the difference in multiple correlations, the interval is

$$0.0166 \leq P_2 - P_1 \leq 0.0452.$$

The interval does not contain zero, which suggests that a significant increment to the validity of prediction can be expected across all populations. (Similarly, one can compute a Z test; for these data the value is $Z = 4.21$, $p < .001$).

The confidence interval for the population difference in squared correlations is

$$0.0137 \leq P_2^2 - P_1^2 \leq 0.0354,$$

and the test of the null hypothesis that the population squared difference equals zero is $Z = 4.44$ ($p < .0001$). Again the results indicate a nonzero incremental validity across schools.

CONCLUSIONS

Pooling estimates across sites provides a viable strategy for estimating the incremental validity. If a single sample is used in each site to assess incremental validity, the test for the statistical significance of the pooled estimate will have adequate power to detect increments in validity of .02 given pooled sample sizes of $N \geq 4{,}000$.

RECOMMENDATION

Estimates of the incremental validity of alternative test batteries should be based on pooled estimates derived from several samples, using the methods outlined in this report.

REFERENCES

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis.* New York: John Wiley.

Becker, B. J. (1985). *Applying tests of combined significance hypotheses and power considerations.* Unpublished doctoral dissertation, The University of Chicago.

Becker, B. J. (1987). Applying tests of combined significance in meta-analysis. *Psychological Bulletin, 102,* 164-171.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions.* Urbana, Illinois: University of Illinois Press.

Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). London: Oliver & Boyd.

George, E. O. (1977). *Combining independent one-sided and two-sided statistical tests: Some theory and applications.* Unpublished doctoral dissertation, Department of Statistics, The University of Rochester.

Hedges, L. V. (1983). Combining independent estimators in research synthesis. *British Journal of Mathematical and Statistical Psychology, 36,* 123-131.

Johnson, N. L., & Kotz, S. (1970). *Distributions in statistics: Continuous univariate distributions-2.* New York: Wiley.

Kozoil, J. A., & Perlman, M. D. (1978). Combining independent chi-squared tests. *Journal of the American Statistical Association, 73,* 753-763.

Lawley, D. (1943). A note on Karl Pearson's selection formulae. *Royal Society of Edinburgh Proceedings, Section A, 62,* 28-30.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Miller, P. G. (1966). *Simultaneous statistical inference* (pp. 67-70). New York: McGraw-Hill.

Mosteller, F., & Bush, R. R. (1954). Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of social psychology: Vol. 1. Theory and method* (pp. 289-334). Cambridge, MA: Addison-Wesley.

Mudholkar, G. S., & George, E. O. (1979). The logit statistic for combining probabilities: An overview. In J. S. Rustagi (Ed.), *Optimizing methods in statistics* (pp. 345-365). New York: Academic Press.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics, 29,* 201-211.

Olkin, I., & Siotani, M. (1976). Asymptotic distribution of functions of a correlation matrix. In S. Ikeda (Ed.), *Essays in probability and statistics* (pp. 235-251). Tokyo: Sinko Tsusho.

Olkin, I., & Sylvan, M. (1977). Correlational analysis when some variances and covariances are known. In P.R. Krishnaiah (Ed.), *Multivariate analysis - IV* (pp. 175-191). Amsterdam: North-Holland.

Oosterhoff, J. (1969). Combination of one-sided statistical tests. *Mathematical Centre Tracts,* 28. Amsterdam: Mathematisch Centrum.

Pearson, K. (1933). On a method of determining whether a sample of size supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika, 25,* 379-410.

Pearson, K., & Filon, L. N. G. (1898). Mathematical contributions to the theory of evolution-IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society, Series A, 191,* 229-311.

Rao, C. R. (1973). *Linear statistical inference* (2nd ed.). New York: John Wiley.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85,* 185-193.

Ryan, B., Joiner, B., & Ryan, T. (1985). *Minitab handbook*. Boston: PWS-Kent.

SAS Institute. (1990). *SAS procedures guide* (Version 6, 3rd ed.). Cary, NC: SAS Institute Inc.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529-540.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, R. M., Jr. (1949). *The American soldier: Adjustment during Army life. (Vol. 1)*. Princeton, NJ: Princeton University Press.

SPSS Inc. (1988). *SPSS-X user's guide* (3rd ed.). Chicago: SPSS Inc.

Tippett, L. H. C. (1931). *The method of statistics*. London: Williams & Norgate.

Wherry, R. J. (1931). A new formula for predicting shrinkage of multiple correlation. *Annals of Mathematical Statistics, 2,* 440-457.

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological Bulletin, 48,* 156-158.

Winer, B. J. (1971). *Statistical principles of experimental design* (2nd ed.). New York: McGraw-Hill.

APPENDIX A

STATISTICAL ANALYSIS SYSTEM (SAS) PROGRAM TO COMPUTE COMBINED SIGNIFICANCE TESTS

SAS Program to Compute Combined Significance Tests

```
OPTIONS NOCENTER;
DATA ONE;
INPUT RS1 RS2 A B N;
DIFF=RS2-RS1;
DF1=B-A;
DF2=N-B-1;
F=DIFF*DF2/(DF1*(1-RS2));
P=1-PROBF(F,DF1,DF2);
Z=PROBIT(P);
LOGP=LOG(P);

CARDS;
.160 .176 10 20 470
.144 .180 10 20 530
.194 .224 10 20 460
.062 .084 10 20 700
;
PROC PRINT; VAR RS1 RS2 N A B F P;
PROC PRINT; VAR DIFF Z LOGP;
PROC MEANS NOPRINT SUM N;
  VAR Z LOGP;
  OUTPUT OUT=SUMS SUM=SUMZ SUML N=K;
DATA TCS; SET SUMS;
ZS=SUMZ/SQRT(K);   CF=-2*SUML;
PS=1-PROBNORM(ZS);   PC=1-PROBCHI(CF,2*K);
ENDSAS;
```

APPENDIX B

A SIMULATION STUDY OF THE DISTRIBUTION OF THE DIFFERENCE IN SQUARED MULTIPLE CORRELATIONS

Page

LIST OF TABLES

Page

A Simulation Study of the Distribution of the Difference in Squared Multiple Correlations[1]

Background

The main body of this report has presented asymptotic formulas for the variance in the difference in multiple correlations or squared multiple correlations, both for the case of independent samples and for the case when the two correlations are based on the same sample but different sets of predictors. The latter case breaks down into three sub-cases, the first two of which are most important:

(a) the second set of predictors includes the first as a subset (IS)

(b) the two sets of predictors are disjoint (DJ)

(c) the two predictor sets overlap but are neither inclusive subsets nor disjoint.

The formulas were intended to apply to several current studies of the incremental validity of adding new aptitude tests to the 10-test Armed Services Vocational Aptitude Battery (ASVAB). In a recent study, Wolfe (1991) reported the validities for predicting school performance in nine Navy technical training schools when four new predictors were added to the ASVAB. Sample sizes ranged from 97 to 929. The validity increments were 0, 0, .001, .007*, .014, .014**, .018**, .029*, and .051**. Subsequent significance tests for the increase in validity from adding a single predictor to the ASVAB showed highly significant improvement for increases as small as .004 when the sample size was 929. The mean validity increase across schools ranged from .002 to .006 when only one predictor was added to the ASVAB. This is an example of Case *a* described in the first paragraph.

In the same study, an alternate form of the ASVAB was re-administered after enlistment, and its validity was compared with the pre-enlistment ASVAB. After correction for range restriction, the two batteries differed by only .009, on the average. This is an example of Case *b* comparison.

Problem

The variance formulas assume asymptotic normality of the difference in squared multiple correlations. But in Case a, the sample difference in squared correlations is non-negative. If the true population difference is zero, the sample differences will approach zero as the sample size increases, while remaining non-negative. Such a distribution cannot be normal. If the population difference is non-zero but small, we can expect slow convergence toward normality as the sample size increases. The rate at which the sample difference approaches normality will determine whether the asymptotic approximations given earlier in this report will have practical utility.

Approach

In order to study the behavior of the asymptotic formulas, simulations were performed with six different sets of artificially specified population parameters and three different sample sizes. Table B-1 shows the characteristics of different samples. The samples for inclusive predictor subsets are labeled with the initial letters IS, and the disjoint sets with the initial letters DJ. The two letters are followed by three digits indicating the true difference in squared multiple correlation. Sample sizes of 100, 400, and 1000 are designated A, B, and C respectively.

For each simulated sample, uniform pseudo-random numbers were generated by a method due to L'Ecuyer (1988). These were converted to a Gaussian (0,1) distribution by a circular transformation described by Knuth (1981, p. 116ff.). Finally, a Cholesky factorization of the population correlation matrix of predictors and criterion was used to generate a multivariate normal distribution of raw scores.

[1]This appendix was written by John H. Wolfe.

* $p < .05$ ** $p < .01$

There were 1000 replications for each sample size and each population correlation matrix. For example, set IS003B consisted of 1000 samples of 400 observations drawn from a population where the true difference in the squared multiple correlation was .003.

Table B-1. Description of Simulation Samples

Samples	N	r_{12}	r_{1y}	r_{2y}	R_1	R_2
		Inclusive Predictor Sets: $\{x_1\}, \{x_1, x_2\}$				
IS000	1000	.50	.400	.200	.4000	.4000
IS003 A,B,C	100, 400, 1000	.30	.500	.200	.5000	.5027
IS006 A,B,C	100, 400, 1000	.60	.400	.300	.4000	.4070
IS013 A,B,C	100, 400, 1000	.50	.400	.300	.4000	.4163
		Disjoint Predictor Sets: $\{x_1\}, \{x_2\}$				
DJ000	100	.99	.400	.400	.4000	.4000
DJ010 A,B,C	100, 400, 1000	.70	.455	.466	.4550	.4660

In each sample, the sample correlation matrix was computed, along with the multiple correlations, their difference, and the difference's asymptotic variance estimated from sample values. These were compared with the asymptotic variance estimated from population values, and with the standard deviation observed across replications.

Results

Table B-2 compares the means and standard deviations of squared correlation differences with their theoretical values. Here δ^* = the population difference in squared multiple correlations and d^* is its sample value. σ_{d^*} is the theoretical asymptotic standard deviation of d^* based on population values (Result 3, Equation 26), and $SDEV_{d^*}$ is the standard deviation observed across the 1000 replications. Note the singularities in the first (IS000) sample, where the population multiple correlations are identical. Asymptotic normal theory breaks down in this case by predicting a zero value for σ_{d^*}. Column six measures the deviation of $\overline{d^*}$ from its theoretical value; the denominator is the theoretical standard error of $\overline{d^*}$ across 1000 replications. If the theory is correct, column six will be a normal (0, 1) deviate.

The sixth column of Table B-2 is a significance test for the difference between d^* and its theoretical value. Except for the DJ010A sets, there is no significant deviation of d^* from its theoretical value. $SDEV_{d^*}$ is compared with its theoretical value in the last two columns of Table B-2. All but the last two inclusive subset samples have significantly greater variance than asymptotic theory predicts. The disjoint sets are in substantial agreement with theory.

Table B-3 displays various measures of normality for d^*. All of the inclusive predictor subsets are non-normal, even for large samples, while all of the disjoint sets are normal.

Table B-4 shows the behavior of $\hat{\sigma}_{d^*}$, the sample estimates of the standard deviation of d^*, computed from Eq. (8) using the replication's sample correlations. Each replication has a different $\hat{\sigma}_{d^*}$. This should be compared with σ_{d^*} in Table B-2, which uses population correlations in a similar formula of Result 3, and with $SDEV_{d^*}$ in Table B-2, which is an observed value across replications. Column 6 shows the correlation between d^* and $\hat{\sigma}_{d^*}^2$. In the usual sampling theory based on normal parent distributions, these would be expected to be independent, not correlated. This independence is essential for using Student's t-distribution to establish confidence intervals. Here the correlations are greater than .99 for all inclusive subsets. (Probably the only reason they are not 1.00 is rounding error in the values of d^* and $\hat{\sigma}_{d^*}$, which were rounded to 4 digits.) The distribution of the T-statistic is shown in the right-most four columns of

Table B-2. Means and Standard Deviations of Squared-Correlation Differences

Sample	N	δ^*	δ^* + Bias	$\overline{d^*}$	$\frac{\overline{d^*}-\delta^*-bias}{\sigma_{d^*}/\sqrt{1000}}$	σ_{d^*}	$SDEV_{d^*}$	$(\frac{SDEV_{d^*}}{\sigma_{d^*}})^2$	$P(\chi^2_{999})$
Inclusive Predictor Sets									
IS000	1000	0.000000	0.000841	0.000900	∞	0.000000	0.001239	∞	0
IS003 A	100	0.002747	0.010280	0.010082	-0.690111	0.009050	0.013669	2.281392	$< 10^{-100}$
IS003 B	400	0.002747	0.004609	0.004767	1.106663	0.004525	0.005339	1.392384	$< 10^{-14}$
IS003 C	1000	0.002747	0.003490	0.003514	0.258851	0.002862	0.003086	1.162848	.0003
IS006 A	100	0.005625	0.013949	0.013412	-1.242703	0.013670	0.017656	1.668217	$< 10^{-35}$
IS006 B	400	0.005625	0.007685	0.007488	-0.910621	0.006835	0.007296	1.139543	.0014
IS006 C	1000	0.005625	0.006447	0.006422	-0.185840	0.004323	0.004555	1.110381	.0083
IS013 A	100	0.013333	0.021404	0.022382	1.480195	0.020880	0.024662	1.395064	$< 10^{-14}$
IS013 B	400	0.013333	0.015330	0.015074	-0.774696	0.010440	0.010393	0.990961	.5744
IS013 C	1000	0.013333	0.014130	0.014181	0.241779	0.006603	0.006455	0.955796	.8384
Disjoint Predictor Sets									
DJ000	100	0.000000	0.000000	-0.000061	-0.185280	0.010360	0.010459	1.019283	.3288
DJ010 A	100	0.010131	0.009914	0.004970	-2.520654	0.062030	0.064624	1.085380	.0307
DJ010 B	400	0.010131	0.010077	0.009863	-0.217976	0.031015	0.029938	0.931754	.9390
DJ010 C	1000	0.010131	0.010109	0.010460	0.565348	0.019616	0.019997	1.039268	.1891

Table B-3. Normality of Squared-Correlation Differences

Sample	N	δ^*	Skewness	Kurtosis	Kolomogorov D	P
Inclusive Predictor Sets						
IS000	1000	0	2.3991	6.9438	.2338	<.01
IS003A	100	.0027	3.0044	16.5396	.2304	<.01
IS003B	400	.0027	1.8945	4.8198	.1860	<.01
IS003C	1000	.0027	1.4189	2.6474	.1274	<.01
IS006A	100	.0056	2.4313	8.1191	.2237	<.01
IS006B	400	.0056	1.7039	4.0798	.1524	<.01
IS006C	1000	.0056	1.0519	1.5267	.0826	<.01
IS013A	100	.0133	1.7557	3.6659	.1821	<.01
IS013B	400	.0133	.9537	1.267	.0735	<.01
IS013C	1000	.0133	.5916	.5317	.0442	<.01
Disjoint Predictor Sets						
DJ000	100	0	.0325	.0058	.2874	>.15
DJ010A	100	.0101	.0872	.3230	.0263	.09
DJ010B	400	.0101	-.0180	.0241	.0197	>.15
DJ010C	1000	.0101	-.0255	.1784	.0197	>.15

Table B-4. Here d^*_w is the value of d^* corrected for bias using the Wherry (1931) shrinkage formula. Normal theory would predict that T would have zero mean, unit standard deviation, and no skewness or kurtosis. The obtained values for disjoint sets are in line with these expectations, but not for the inclusive subsets.

Table B-4. Estimated Errors and T-ratios for Squared-Correlation Differences									
Sample	N	δ^*	$\hat{\sigma}_{d^*}$	$\hat{\sigma}_{d^*}$	$r(d^*, \hat{\sigma}_{d^*}^2)$	$T = (\delta^* - d_w^*)/\hat{\sigma}_{d^*}$			
			Mean	Std. Dev.		Mean	Std. Dev.	Skewness	Kurtosis
Inclusive Predictor Sets									
IS000	1000	.0000	.001395	.001034	.9992	1.5284	5.4964	8.4747	95.7577
IS003A	100	.0027	.013778	.009907	.9914	2.2943	9.7766	9.6427	111.3506
IS003B	400	.0027	.004995	.003169	.9979	2.4773	15.4916	16.10133	297.4342
IS003C	1000	.0027	.002896	.001430	.9989	1.4652	6.1591	8.0970	83.2858
IS006A	100	.0056	.016744	.011741	.9922	3.3358	15.0184	8.9495	95.4991
IS006B	400	.0056	.006870	.003715	.9982	1.9732	12.0017	15.0540	266.6279
IS006C	1000	.0056	.004287	.001667	.9990	.8876	4.9319	14.4417	259.7843
IS013A	100	.0133	.022233	.013608	.9920	4.2551	32.4311	15.5740	289.0354
IS013B	400	.0133	.010305	.003907	.9971	.8270	4.2313	13.2275	218.1110
IS013C	1000	.0133	.006601	.001565	.9979	.2781	1.2850	2.1404	9.6911
Disjoint Predictor Sets									
DJ000	100	.0000	.010196	.001742	-.0408	-.0002	.9976	-.0262	-.2136
DJ010A	100	.0101	.060657	.006639	.0274	.0888	1.0672	-.0544	.1019
DJ010B	400	.0101	.030891	.001664	.0018	.0070	.9685	.0098	-.0170
DJ010C	1000	.0101	.019588	.000657	.0067	-.0185	1.0205	.0422	.1598

Finally, Table B-5 shows what would happen if one tried to base confidence intervals on the asymptotic estimates of variance. The middle three columns show the 5 percent, the median, and the 95 percent values of the T-statistic observed among the 1000 replications of a sample. Normal theory would predict these values to be -1.645, 0, and +1.645. The disjoint predictor samples come close, but the inclusive predictor subsets do not. The last two columns of Table B-5 show the number of replications in which δ^* falls outside of a "confidence interval" of $d^* \pm 1.96\hat{\sigma}_{d^*}$. Normal theory would expect 25±10 at each end. The observed frequencies for the disjoint sets are close, but the inclusive subsets are grossly deviant from normal theory.

Discussion

The asymptotic formulas derived in the main body of this report seem to work very well when the predictor sets are disjoint, but are less satisfactory, even on large samples, when one set includes another. Several alternative remedies could be tried:

Use the mean squared error (MSE) instead of multiple correlation. Sympson (1979) suggested this approach because, when adjusted for degrees of freedom, the difference in MSEs can be negative. However, the difference in such MSEs is proportional to the difference in Wherry-corrected squared multiple correlations (Wherry, 1931). Although not shown in the tables, the Wherry-corrected values had skewness and kurtosis values that agreed with the uncorrected values to two decimals in most of these simulations.

Transform the multiple correlations. A Fisher z-transform or other transform will still allow the difference to approach its population value with increasing sample size, while remaining non-negative. If the population difference is zero, the distribution of the difference cannot be normal.

Transform the difference in squared multiple correlations. Since the distributions are sometimes almost J-shaped, they are difficult to normalize. It would be desirable if a variance-stabilizing transformation could be found to eliminate the large correlation between d^* and $\hat{\sigma}_{d^*}^2$, as well as normalize d^*.

Table B-5. Confidence Intervals for Squared-Correlation Differences

			$T = (\delta^* - d_w^*)/\hat{\sigma}_{d^*}$			Frequenc	
Sample	N	δ^*	5%	50%	95%	$\delta^* < d_w^* - 1.96\hat{\sigma}_{d^*}$	$\delta^* > d_w^* + 1.96\hat{\sigma}_{d^*}$
Inclusive Predictor Sets							
IS000	1000	.0000	-.7699	.3616	6.2172	6	171
IS003A	100	.0027	-.8783	.4119	7.7085	1	21
IS003B	400	.0027	-1.0196	.3106	8.0732	2	227
IS003C	1000	.0027	-1.1226	.2455	6.3724	2	159
IS006A	100	.0056	-.9414	.4984	10.0743	3	228
IS006B	400	.0056	-1.0896	.3175	6.8493	2	199
IS006C	1000	.0056	-1.1728	.2011	3.6934	4	141
IS013A	100	.0133	-1.1432	.3493	10.1649	5	207
IS013B	400	.0133	-1.1508	.1914	3.8499	3	137
IS013C	1000	.0133	-1.2847	.0838	2.3916	5	78
Disjoint Predictor Sets							
DJ000	100	.0000	-1.6850	-.0253	1.6430	21	19
DJ010A	100	.0101	-1.7213	.1089	1.9215	31	48
DJ010B	400	.0101	-1.6067	.0261	1.6441	22	26
DJ010C	1000	.0101	-1.6524	-.0467	1.6052	30	26

Use resampling techniques. One could, of course, abandon attempts to obtain explicit mathematical expressions for the sample variance of multiple correlation differences and use resampling techniques to estimate variances in each data sample. While it may be useful in practice, such an approach is beyond the scope of this paper.

Use other sampling distributions. The inadequacy of the above approaches almost forces us to use non-Gaussian sampling distributions. These are outlined in the next section.

Non-normal Sampling Theory for Correlation Differences

Let $\delta = \frac{\delta^*}{1-P_Y^2}$ with sample estimate $\tilde{d} = \frac{d^*}{1-R_Y^2}$. ($\delta$ is often called the *effect size.*) Let $u = b - a$ and $v = N - b - 1$ be the degrees of freedom. When $\delta^* = 0$, the distribution of $(\frac{v}{u})\tilde{d}$ is exactly central $F_{u,v}$. When $\delta^* > 0$, the distribution is a non-central F with non-centrality parameter $\lambda = (N-a)\delta$ (Cohen, 1988, p. 551).[2] The mean of F is given by

$$\bar{F} = \frac{v}{v-2}(1+\frac{\lambda}{u}). \tag{B-1}$$

From this, it is readily seen that an unbiased estimate of δ is

$$\hat{\delta} = \frac{b-a}{N-a}(\frac{N-b-3}{N-b-1}F - 1). \tag{B-2}$$

Or, in terms of $\tilde{d}$, an unbiased estimate of δ is

$$\hat{\delta} = \frac{(N-b-3)\tilde{d} - b + a}{N-a}. \tag{B-3}$$

[2]Lee(1971) has developed more accurate non-central F approximations for the special case when a = 0 by fitting the first three moments.

For N > 100, the non-central χ^2 will serve nearly as well. Here, $\chi^2_u = v\bar{d}$ with the same non-centrality parameter λ. The mean of a non-central χ^2 is given by

$$\overline{\chi^2_u} = u + \lambda. \tag{B-4}$$

Then the point estimate for δ is

$$\hat{\delta} = \frac{\chi^2 - b + a}{N - a}. \tag{B-5}$$

Or,

$$\hat{\delta} = \frac{(N-b-1)\bar{d} - b + a}{N - a}. \tag{B-6}$$

In comparing this formula with the one for the non-central F, it is seen that the χ^2-based estimate of δ is biased upward by $2\bar{d}/(N-a)$. However, it is easier to average results across samples and compute confidence intervals with the non-central χ^2 distribution than with the non-central F.

Suppose that there are k such χ^2 populations with possibly different values of δ. Then the sum

$$S = \sum_{i=1}^{k} (N_i - b - 1)\bar{d}_i \tag{B-7}$$

has a non-central χ^2 distribution with ku degrees of freedom and non-centrality parameter

$$\lambda = \sum_{i=1}^{k} (N_i - a)\delta_i. \tag{B-8}$$

A weighted mean of the δ_i across the different populations can be defined as $\delta = \frac{\lambda}{N - ka}$, where $N = \sum_{i=1}^{k} N_i$. Then a nearly unbiased estimate of $\delta = Mean(\hat{\delta}_i)$. Hence,

$$\hat{\delta} = \frac{S - k(b-a)}{N - ka}. \tag{B-9}$$

The value of S can be used to compute the upper and lower 2.5 percent limits for a confidence interval $\lambda_{.025} \le \lambda \le \lambda_{.975}$ by means of Applied Statistics algorithm 170 (Narula & Desu, 1981). From this, it is evident that a 95 percent confidence interval can be established around the weighted mean of δ with

$$\frac{\lambda_{.025}}{N - ka} \le \delta \le \frac{\lambda_{.975}}{N - ka}. \tag{B-10}$$

There is some difficulty in relating these results to d^* itself. Since $\bar{d}$ is a function of both $R^2_{\hat{Y}}$ and d^*, the relation between $\bar{d}$ and d^* is not one to one. The transformation $w = -log(1-R^2)$ may be useful here.[3] Differentiating w (or expanding it in a Taylor series) shows that, to a first order approximation, $\Delta w \equiv w_2 - w_1 = \bar{d}$. Thus it may be possible to develop all of the results in the w metric rather than the R metric.

Simulation Results with Non-normal Sampling Distributions

Tables B-6 and B-7 compare the observed and theoretical values of the first four moments of the non-central F model of $\bar{d}$ and the non-central χ^2 model of Δw, respectively. It is evident that both models fit the first three moments rather well, and that the non-central F model fits the fourth moment better than the non-central χ^2. Results for the non-central χ^2 model of $\bar{d}$ are not shown, but were were no better than those shown in Table B-7; in fact the ratios of observed to theoretical variances deviated slightly more from 1 than they did in Table B-7.

[3] It is interesting to note that Moschopoulos (1983) has shown that w, raised to a suitable power is approximately Gaussian.

Table B-6. Observed vs. Theoretical Moments of Non-Central F fitted to $\frac{d^*}{1-R_Y^2}$

			Mean		Variance			Skewness		Kurtosis	
Sample	N	δ^*	Sample	Theory	Sample	Theory	Ratio	Sample	Theory	Sample	Theory
IS000	1000	0.	0.96	1.00	2.02	2.01	1.01	2.75	2.84	9.66	12.15
IS003A	100	0.0027	1.43	1.39	3.91	3.72	1.05	2.54	2.74	8.20	11.32
IS003B	400	0.0027	2.56	2.48	8.51	8.02	1.06	1.86	1.99	3.95	5.56
IS003C	1000	0.0027	4.86	4.68	17.48	16.84	1.04	1.31	1.42	2.31	2.77
IS006A	100	0.0056	1.86	1.70	6.00	5.04	1.19	2.33	2.51	6.49	9.37
IS006B	400	0.0056	3.69	3.71	11.88	13.02	0.91	1.49	1.62	3.36	3.66
IS006C	1000	0.0056	7.82	7.75	29.89	29.24	1.02	1.06	1.10	1.28	1.66
IS013A	100	0.0133	2.58	2.65	8.41	9.08	0.93	1.87	2.04	4.32	6.06
IS013B	400	0.0133	7.49	7.47	28.74	28.45	1.01	1.13	1.15	1.80	1.83
IS013C	1000	0.0133	17.35	17.15	70.03	67.45	1.04	0.70	0.75	0.52	0.77

Table B-7. Observed vs. Theoretical Moments of Non-Central χ^2 fitted to $\Delta\log(1-R^2)$

			Mean		Variance			Skewness		Kurtosis	
Sample	N	δ^*	Sample	Theory	Sample	Theory	Ratio	Sample	Theory	Sample	Theory
IS000	1000	0.	0.96	1.00	2.01	2.00	1.01	2.74	2.83	9.59	10.00
IS003A	100	0.0027	1.40	1.36	3.62	3.46	1.05	2.43	2.60	7.43	7.87
IS003B	400	0.0027	2.54	2.47	8.29	7.87	1.05	1.83	1.96	3.80	3.32
IS003C	1000	0.0027	4.84	4.68	17.22	16.70	1.03	1.30	1.41	2.22	0.70
IS006A	100	0.0056	1.82	1.67	5.47	4.66	1.17	2.23	2.38	5.79	6.09
IS006B	400	0.0056	3.66	3.68	11.51	12.72	0.90	1.45	1.59	3.12	1.48
IS006C	1000	0.0056	7.78	7.71	29.26	28.85	1.01	1.04	1.09	1.22	-0.39
IS013A	100	0.0133	2.50	2.58	7.58	8.34	0.91	1.76	1.91	3.68	3.07
IS013B	400	0.0133	7.39	7.38	27.29	27.54	0.99	1.08	1.12	1.59	-0.32
IS013C	1000	0.0133	17.17	16.98	67.28	65.94	1.02	0.68	0.73	0.46	-1.28

Applied Statistics algorithm 170 (Narula & Desu, 1981) for the non-central χ^2 was used to compute 95 percent confidence intervals based on the sample value of $\bar{d}$ in each replication. The numbers of replications (out of a 1000) for which the population values of δ or Δw fell above or below the confidence intervals were tabulated in Table B-8. Applying the binomial distribution to the frequencies, (N = 1000, p = .025), the observed frequency should lie in the range 25±10 for 95.8 percent of the ten simulated populations at the upper end and 95.8 percent of the ten populations at the lower end. Frequencies that lie outside the range 25±10 are marked with an asterisk.

For the χ^2 model of $\bar{d}$, two of the ten populations had 95 percent confidence intervals that were too often above δ and three of them were too often low. For the χ^2 model of Δw, none of the ten populations had 95 percent confidence intervals that were too often above the population value of Δw and three of them were too often low.

When $\delta^* = 0$, there were no sample estimates of the upper 2.5 percent limit that were less than the true value, zero. Every sample estimate of the lower 2.5 percent limit was greater than zero, usually in the

Table B-8. Frequency of Samples with Population Values Falling Outside 95% Confidence Intervals based on Non-central χ^2 Models

			$\frac{d^*}{1-R_Z^2}$		$\Delta log\,(1-R^2)$	
Sample	N	δ^*	N Below 2.5% Limit	N Above 97.5% Limit	N Below 2.5% Limit	N Above 97.5% Limit
IS000	1000	0.	30	0	30	0*
IS003A	100	0.0027	27	27	27	27
IS003B	400	0.0027	36*	32	33	32
IS003C	1000	0.0027	27	21	25	21
IS006A	100	0.0056	41*	13*	34	13*
IS006B	400	0.0056	21	29	19	29
IS006C	1000	0.0056	23	22	23	22
IS013A	100	0.0133	27	39*	25	39*
IS013B	400	0.0133	25	25	23	25
IS013C	1000	0.0133	28	21	26	21

eighth decimal place. This latter effect was corrected by subtracting 10^{-6} from the lower 2.5 percent limits computed by the program.

Although the χ^2 approximations for the confidence intervals do not perform as well as might be expected for non-central F formulas in these simulations, they work quite well for the largest sample sizes. In any case, they are a substantial improvement over the normality-based confidence intervals in Table B-5.

Example

Table B-9 shows the application of these methods to the illustrative data in Tables 7, 10, and 11 in the main body of the report. The confidence intervals and unbiased estimates are based on the non-central chi-square distribution. The p_i values at the right-most column are based on the *central* chi-square distribution, and are close to the F probabilities in Table 7. The probability for the combined sample, .004, agrees with the Fisher value previously given.

Table B-9. Ninety -Five Percent Confidence Intervals for Effect Sizes

School	Effect Size $\frac{d^*}{1-R_Z^2}$	Unbiased Effect $\hat{\delta}$	Lower Limit 2.5^{σ}	Upper Limit 97.5%	p_i
Air Traffic Controller	.0199	-.0024	.0000	.0275	.538
Fire Control Technician	.0431	.0228	.0014	.0642	.016
Gunner's Mate	.0388	.0236	.0045	.0575	.003
Electrician's Mate	.0236	.0007	.0000	.0337	.410
Combined Sample	.0319	.0130	.0027	.0283	.004

Conclusion

Asymptotic normal theory works well when the predictor sets are disjoint. When one predictor set includes another, then non-central F or chi-square distributions may be used to establish confidence intervals for the effect sizes in each sample, and for the mean effect size across different samples. Unfortunately, it is not clear how to test hypotheses concerning differences among populations. Adjusting effect sizes for artifacts of range restriction or criterion unreliability will require further research.

REFERENCES

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* Hillsdale, NJ: Lawrence Erlbaum Associates.

Knuth, D. (1981). *Seminumerical algorithms (2nd ed.),* vol. 2 of *The art of computer programming.* Reading, Mass: Addison-Wesley. pp. 116ff.

L'Ecuyer, P. (1988). Efficient and portable combined random number generators. *Communications of the ACM, 31,* 742-749, 774.

Lee, Y. (1971). Some results on the sampling distribution of the multiple correlation coefficient. *Journal of the Royal Statistical Society, Series B, 33,* 117-130.

Moschopoulos, P. G. (1983). On a new transformation to normality. *Communications in Statistics - Theory and Methods, 12,* 1873-1878.

Narula, S., & Desu, M. (1981). Computation of probability and non-centrality parameter of a non-central chi-squared distribution. *Applied Statistics, 30,* 349-352.

Sympson, J. (1979). *Testing differences between multiple correlations* (Research Report RR-79-20). Princeton, NJ: Educational Testing Service.

Wherry, R. J. (1931). A new formula for predicting shrinkage of multiple correlation. *Annals of Mathematical Statistics, 2,* 440-457.

Wolfe, J. (1991, July). *Navy validity study of new predictors: Preliminary findings.* Briefing for the Defense Advisory Committee on Military Personnel Testing, Monterey, CA.

Distribution List

Office of the Assistant Secretary of Defense (FM&P) (2)
HQUSMEPCOM (MEPCT-P)
Director, Research and Development Department of Defense Coordinator (PERS-234) (3)
Defense Technical Information Center (DTIC) (4)

Copy to:
Deputy Chief of Naval Operations (MP&T) (N1)
Deputy Chief of Naval Personnel (N1B)
Director, Recruiting and Retention Programs Division (PERS-23)
Assistant for Planning and Technical Development (PERS-01JJ)
Office of the Director, Test & Evaluation and Technology Requirements (N091)
Office of Chief of Naval Research (ONT-222)
Army Research Institute, AISTA (PERI II)
AL/DOKLO Technical Library, Brcoks AFB, TX
Armstrong Laboratory, Operations & Support Directorate, Technical Information Services Division (STINFO), Brooks AFB, TX
Director of Research, U.S. Naval Academy
Center for Naval Analyses (Acquisitions Unit)
Center for Naval Analyses
Systems Research Center, Virginia Tech., Blacksburg, VA